
Human Resource
Management
in the Era of Big Data
大数据时代的
人力资源管理
蔡治◎著

清华大学出版社

大数据时代的人力资源管理

蔡　治　著

清华大学出版社

北　京

内容简介

本书采取人物对话的形式，用讲故事的方法，将人力资源管理中一些典型的问题用高级数据分析的方法去解决。

全书分为 8 章，第 1～2 章介绍人力资源管理数据分析的意义和数据分析前的准备工作；第 3 章讲述回归分析法在员工需求预测中的应用；第 4 章讲述培训师评估分数的标准化；第 5 章分析薪酬公平性；第 6 章介绍综合评价法在员工能力评估中的应用；第 7 章介绍如何使用 Boosting、随机森林算法预测员工离职概率；第 8 章讲述如何通过文本分析中的情感分析法解读员工辞职报告。

本书能够帮助人力资源管理人员开阔眼界、打开思维，加深对数据分析的认识，促进数据分析技术在人力资源管理领域的应用。

图书在版编目(CIP)数据

大数据时代的人力资源管理/蔡治著. --北京：清华大学出版社，2016
ISBN 978-7-302-45089-4

Ⅰ. ①大… Ⅱ. ①蔡… Ⅲ. ①人力资源管理 Ⅳ. ①F243

中国版本图书馆 CIP 数据核字(2016)第 225962 号

责任编辑：刘志彬　张　伟
封面设计：汉风唐韵
责任校对：王荣静
责任印制：王静怡

出版发行：清华大学出版社
网　址：http://www.tup.com.cn，http://www.wqbook.com
地　址：北京清华大学学研大厦 A 座　　**邮　编**：100084
社 总 机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　张**：14.75　　**字　数**：186 千字
版　　次：2016 年 11 月第 1 版　　**印　次**：2016 年 11 月第 1 次印刷
定　　价：39.80 元

产品编号：069752-01

前　言

笔者一直想将概率统计、数据挖掘等数据分析的高级方法应用到人力资源管理领域。在当前的信息化、数据化时代，人力资源管理对数据的依赖性相当强，从招聘中的能力和素质测评，到培训评估、绩效管理、岗位分析、劳动用工、效能分析、薪酬管理等各方面都需要进行数据分析。但人力资源的数据分析大多是描述性统计分析，较少用到高级数据分析技术，如回归分析、聚类分析、因子分析、判别分析、文本挖掘等，对数据的利用率不高，更缺乏对数据的有效和深入挖掘。

笔者一直苦于没有找到合适的工具，直到接触 R 语言。随着了解不断深入，笔者发现 R 语言有很多优点：它摆脱了 SPSS 这类软件的禁锢，即摆脱那种严格的环境和刻板的分析；函数式的编程风格很接近 Excel 函数用法，复杂的模型通常一两个函数就能解决，容易学习和上手；拥有大量的统计算法，可以任意研究和使用；可以绘制出生动美观的数据图形。而且 R 语言完全免费，这对人力资源管理专业人员来说非常重要，因为企业几乎不太可能为人力资源部门专门配备商业统计软件。

于是本书做了一次大胆尝试，即以 R 语言为基础，将概率统计、机器学习、文本挖掘等大数据时代流行的数据分析技术，和人力资源管理实践结合在一起，看看有何化学反应。在此之前，鲜见人力资源管理专业人员涉足这个领域，在此之后，你会发现原来人力资源管理也可以运用大数据分析技术，也可以通过数据挖掘来发现数据价值，也能用机器学习的算法预测未来可能发生的事件，还能对文字内容进行数据分析，而这一切在 R 语言的驱动下变得容易实现。

本书的每个案例都以人力资源管理中的现实情景为基础，通过人物对话的方式来讲述。书中虚拟的谦多顺公司在人力资源管理方面出现了一些问题，比如员工需求数量不准确、员工薪酬满意度不高、学员对培训师的意见比较大、新员工离职率比较高、员工能力评价不够客观、离职沟通出现问题，等等。人力资源部经理 Miss 陈面对这些问题，采用数据分析的方法，帮助部门同事逐个解决问题。在这个过程中，你可以了解概率统计的基本知识、数据挖掘的经典算法，以及文本挖掘中的情感分析，并领略 R 语言的魅力。

本书由于涉及统计学领域的知识，还涉及 R 语言编程，对人力资源管理专业人员来说有一定难度。为此笔者对书中内容做了一些特别设计，比如必须讲的统计知识尽量详细并且图文并茂，所有案例都提供 R 源代码以方便练习，等等。如果潜心阅读，并辅以实践演练，相信会有莫大收获。

希望本书的出版，能够让越来越多的人力资源管理专业人士认识 R 语言，运用高级数据分析技术来有效解决企业中的管理问题，更好地发挥人力资源数据的价值。

为什么编写本书

人力资源管理源于数据分析。20 世纪初古典管理学家弗雷德里克·温斯洛·泰勒通过实验研究如何提高工人的劳动生产率，并提出了

迄今仍在使用的计件工资制、计时工资制，可算作人力资源数据分析的先驱。后来闵斯特伯格、梅奥两位学者将心理学方法引入工业领域，通过大量实验，研究如何提高工人效率，其核心依然是对数据的测量和分析。所以，人力资源管理从发展之初就与数据分析结下不解之缘。一百多年后的今天，世界进入了信息化、数据化时代，但我国人力资源管理却在数据分析领域原地踏步，在大数据门外驻足不前，仍然在汇总、平均、同比、环比，仍然在依赖 Excel，几乎没有将数据挖掘等高级技术应用到管理实践中，去更充分地挖掘数据的价值。这不能不说是一种遗憾！

人力资源管理领域未及时享用数据分析技术发展带来的福利，像那些重要且经典的算法如判别分析、机器学习、聚类分析、因子分析、时间序列分析、文本挖掘等早已进入零售、金融、通信、电子商务以及社交媒体行业，并且表现出令人惊讶的作用，但始终把人力资源管理挡在门外。

然而，人力资源管理专业人员学习数据分析的意愿并不十分强烈。根据弗鲁姆的理论，人力资源管理专业人员研究数据分析的动机强弱，取决于数据分析能够为工作带来的价值大小、学习的难度大小，以及学习的工具和环境的适宜程度。可想而知，在看不到数据分析带来的价值，对数据分析知识心存畏难，且没有称手的分析工具时，人力资源管理专业人员又怎能迈入数据分析的世界呢？

所以，本书尝试将数据分析的高级技术引入人力资源管理领域，提升人力资源管理专业人员学习数据分析的动机水平。首先，用人力资源管理专业人员熟悉的情景编写案例，让大家了解数据分析技术在人力资源管理过程中的作用和价值；其次，穿插普及数据分析的基础知识和算法，重点介绍当前数据分析领域表现优异的统计工具——R 语言，并附送源代码。希望能够唤起看到本书的人力资源管理同行对高级数据分析的兴趣。

当然，本书只是抛砖引玉。鉴于笔者视野狭窄，狭隘地认为我国人力

资源管理领域并未真正涉足数据分析，并未有“大牛”出现，实际上这可能是错误的。不排除有“牛人”早已进行深入的研究，程度之深，应用范围之广，超出笔者的想象。若能发现同行在做同样的事情，希望能够交流、学习，共同促进和提升。

也希望通过本书能够进一步推广R语言。笔者用过不少统计软件，但从未有一款如R语言那样让笔者着迷，它几乎能满足笔者对数据分析的所有需求，分析过程简单快速，各种算法随手拈来，图形绘制变化万千。这么好的统计工具，还是免费的，实在没有理由拒绝，也希望更多的人能够知道这个工具，早早用上。

本书特点

(1) 创新性强，内容为人力资源管理、数据分析和R语言的交叉知识领域。国内首次以R语言为工具，将数据挖掘、文本挖掘等数据分析技术引入人力资源管理领域。

(2) 深入浅出、通俗易懂。全书以人力资源管理人员(简称HR)的视角为基础，采取人物对话方式，结合案例讲解数据分析技术在人力资源管理实践中的应用。

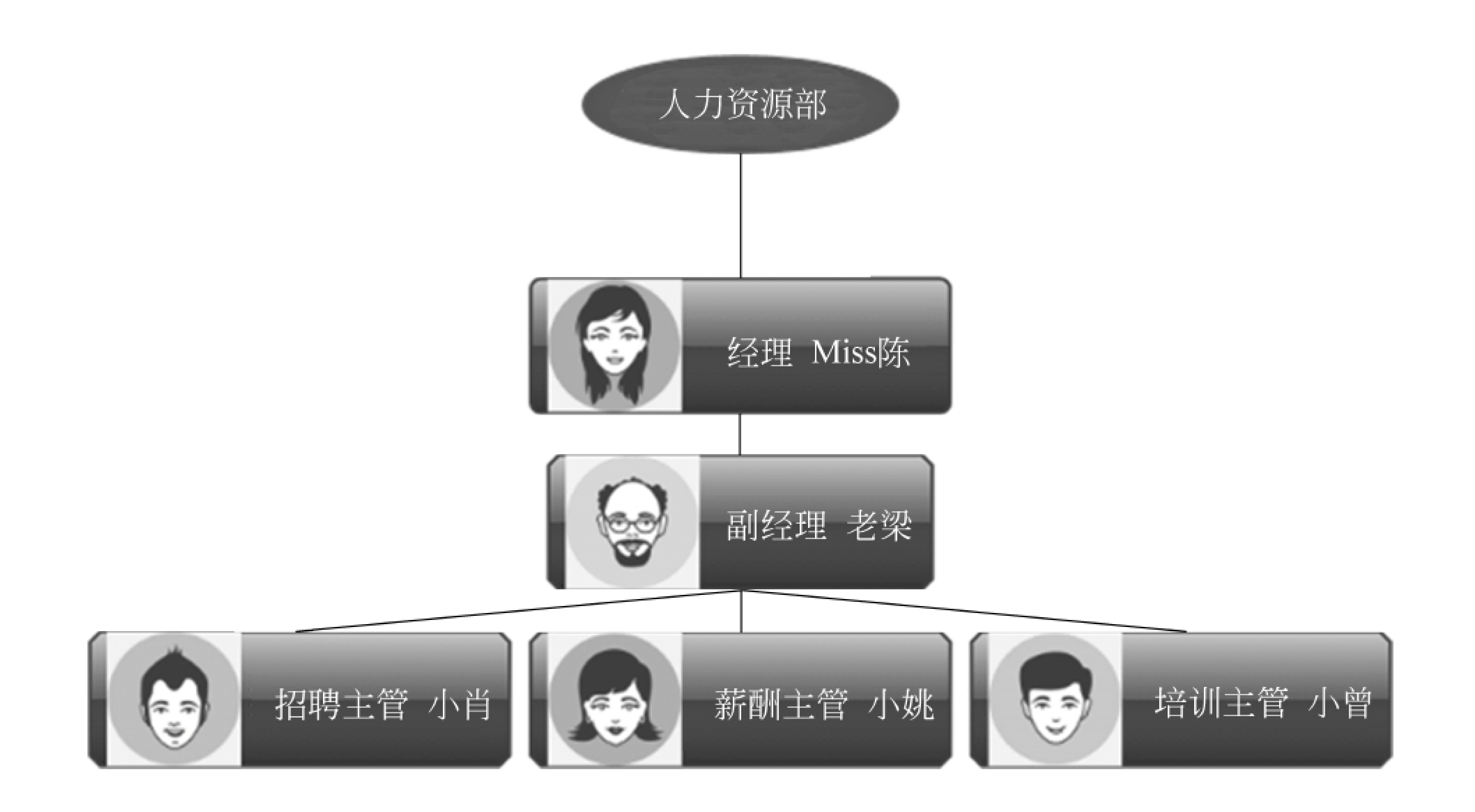
(3) 对HR来说熟悉度高，代入感强，认同感强。书中案例均以人力资源管理中的常见情景为基础，涉及招聘、培训、薪酬、员工关系管理等模块，对HR来说接受程度高。

(4) 阅读难度较低。全书避开讲解复杂的统计学概念、算法，避开讲解R语言的数据结构、语法等内容，重点介绍统计方法的应用案例及其效果，降低阅读难度。

(5) 提供完整源代码和数据。源代码重复使用性高，可直接运行并显示效果，易于操练，方便解读，源代码经小量修改后即可用于各类企业。

本书人物关系图和公司设定

1. 人物关系图



2. 公司设定

公司名称：谦多顺集团股份有限公司

公司规模：下属 20 家子公司，员工 3 万余人

公司性质：民营企业

主营业务：房地产、软件开发、物业服务、通信产品生产与销售等业务。

本书内容

全书共分 8 章，各章内容如下。

第 1 章：人力资源数据分析的意义。介绍人力资源数据分析的特点、难点以及人力资源数据分析和大数据的关系。

第 2 章：数据分析前的准备工作。包括如何选用数据分析的工具，数据收集的工具和方法，以及如何整理数据。

第 3 章：员工年度需求预测。主要介绍了需求预测所采用的方法并分析整个过程。

第 4 章：培训师评估。介绍如何建立企业内部培训讲师授课评分数据库，在此基础上通过计算机标准分建立常模，绘制正态分布图，用定量化的方法选择讲师，并进行培训评估。

第 5 章：薪酬公平性分析。讲解如何运用薪资结构图、基尼系数、Compa 指标、薪酬公平感计量模型来分析员工薪酬公平性。

第 6 章：员工综合能力评估。讲解通过综合评价法评估员工综合能力。

第 7 章：员工离职倾向分析。介绍了如何用 Boosting、随机森林等机器学习算法预测员工未来一年内的离职概率。

第 8 章：员工辞职报告的情感分析。介绍用文本挖掘中的情感分析技术分析员工辞职报告。

关于作者

蔡治：西南师范大学心理学硕士、高级经济师、高级人力资源管理师、高级企业培训师、SPSS 数据分析师，R 语言爱好者，长期从事人力资源管理工作，现任某国有通信企业人力资源部经理。

哪些人会对本书有阅读兴趣

- 人力资源管理工作中需要进行数据分析的人士。
- R 语言爱好者，对 R 语言在各行业中的应用感兴趣的人士。
- 经常阅读分析报告，关注各职能板块研究报告的各级管理人员。

- 从事咨询、研究、分析等工作的专业人士。
- 人力资源管理专业的本科生和研究生。

致谢

感谢广东省通信产业服务有限公司陈洪先生、钟永健先生、冯丽芳女士和张晓军女士，将数据分析的任务交给我，为我提供了在工作中研究和应用数据分析的机会，促成我去接触和学习 R 语言。感谢李延华、张宝、张静，我们经常在一起沟通、讨论，产生了不少想法。感谢夫人陈丽君女士的默默支持和鼓励，让我得以完成本书的写作。

尽管我对书稿校正多次，但仍然不可避免有疏漏和不足之处，请读者批评指正。我会在适当的时间进行修正，以满足大家的需要。

与作者联系

博客：http://blog.sina.com.cn/editcai

邮箱：cizimail@qq.com

作　者

2016 年 8 月

目　录

第 1 章

人力资源数据分析的意义

导语：对人力资源管理专业人员来说，数据分析是一门新技能，而学习这种新技能需要投入成本，包括时间成本、资金成本等。既然要投入成本，自然希望获得回报，并且明白获得回报的难度。按照弗洛姆的期望理论，这两个因素结合在一起才能产生学习动机。本章围绕这两个因素，阐述人力资源管理专业人员为什么需要学习数据分析，学习获得的回报是什么，学习的难度又如何。

1.1 人力资源管理为何需要数据分析

老梁：经理，您常说人力资源管理要重视数据分析，可我觉得人力资源管理在实际工作中并不缺少数据分析啊。您看我们做薪酬、管绩效、建档案、搞培训都是在和数据打交道，每月、每季、每年都会出分析报表，这些不就是数据分析吗？咱们已经在做了，为什么您还强调数据分析呢？

Miss 陈：你说的这些工作自然是在和数据打交道，也是数据分析，但主要是对人力资源各个管理模块产生的数据进行简单的分析运算，如汇总、计算均值、总和等，再通过横向对比、纵向对比等方法从不同维度进行比较分析，然后形成报表，做成报告。实际上，这些工作属于数据分析的较浅层次。

老梁：较浅层次？您的意思是人力资源管理数据分析还分层次吗？

Miss 陈：是的，数据分析的层次和我们人力资源管理的发展阶段有关系，你知道人力资源管理发展的三个阶段吗？

老梁：知道，人力资源管理历经了三个阶段，分别是人事管理阶段、单向人力资源管理阶段和战略人力资源管理阶段。

Miss 陈：其实不同管理阶段对数据分析的需求不同，人力资源管理发展的三个阶段分别对应了三个层次的数据分析需求，具体来说有以下三点。

(1) 人事管理阶段：这个阶段需要对基本数据进行整理、统计，比如计算薪酬、记录考勤、统计加班信息、分类统计人员信息、编制薪资报表等，基本上就是对原始数据进行普通预算，这属于数据粗加工。

(2) 单向人力资源管理阶段：这个阶段在对数据粗加工的基础上，需

要统计更为复杂的指标，用于分析和反映人力资源管理的水平，诊断管理的健康程度。这些指标涉及人力资源各个模块，比如招聘成功率、员工流动率、培训百分比、工作负荷率、企业年轻化程度、劳动生产率，等等。经过几十年的发展，人们总结了不少指标，从类别上划分，大致可以分为人力资源效率指标、人力资源发展指标、人力资源描述指标、人力资源健康指标四类，还形成了人力资源统计学、人力资源会计学等学科。这个阶段开始对数据进行精加工，主要是研究和提炼管理指标，通过计算各种指标来进行数据分析。

(3) 战略人力资源管理阶段：这个阶段将人力资源效能与公司发展战略结合起来，形成人力资源发展战略，进入战略管理阶段。这个阶段需要分析人力资本的投入和回报、人力资源在企业的影响力、人力资源如何促进公司战略目标的实现等更高层次的命题。这个层次需要更为复杂的统计指标和分析技术，在分析指标上重点研究人力资本在企业中发挥的作用，并能够根据需要建立管理分析模型，在分析技术上需要采用更为高级的概率统计分析方法。

老梁：原来不同的发展阶段对人力资源数据分析的需求是不同的，看来我对数据分析的理解还不够啊！

Miss 陈：所以我们也要与时俱进，结合当前人力资源管理的发展趋势，加强对数据分析知识、技能、工具的学习，提高数据分析水平，将数据分析的知识和技术应用到人力资源管理实践中去，提升我们的管理水平，促进公司战略目标的实现。

老梁：经理，您说得对，不过关于数据分析对人力资源管理工作的必要性，您能讲得再详细点吗？咱也想加深对数据分析的认识和理解。

Miss 陈：好的，下面我就详细讲一下人力资源数据分析的意义。

1.1.1 数据分析是人力资源管理发展的趋势

Miss 陈：老梁，请问你现在的工作可以不用电脑吗？

老梁：经理，根本离不开电脑啊。不仅是我，几乎每个部门每个员工的工作都离不开电脑。上个月初公司停了一天电，结果各个部门的工作都停滞了，台式电脑开不了机，内部服务器瘫痪，笔记本电脑即使能用也打不开 OA(办公自动化)。于是大家休息了一天，啥工作都没干成。

Miss 陈：这说明我们的工作对电脑的依赖性很强，超过了以往任何时候。我们已经习惯了通过办公软件和各种管理系统来开展工作。比如，在人力资源管理方面，我们就启用了若干信息化系统来辅助管理，包括员工档案管理系统、培训管理系统、在线培训系统、员工素质测评系统、绩效考核系统等。我们对这些管理系统产生了依赖性，而这种依赖性实际上也成为了当前人力资源管理的特征，照目前的趋势来看，这些管理系统还会逐步向移动终端发展。

计算机管理系统每天都会产生大量数据，如何充分利用这些数据来提升人力资源管理水平，已成为人力资源管理的重要课题。这些数据就像是原材料，我们现在只是进行了粗加工，实际上可以进行精加工，可以更加有效地利用这些数据来为我们所用，给我们提供更有价值的信息。

现代计算机技术的发展、大数据技术的发展、数据挖掘技术的发展，以及数据分析工具的普及，都为高级数据分析技术在人力资源管理领域的应用提供了良好的土壤，也对人力资源管理工作提出了更高的要求。那些看上去复杂、神秘的数据分析技术和昂贵的数据分析软件曾经阻碍了数据分析技术在管理领域的广泛应用，但是现在形势已经发生变化，数据分析的技术和工具不再是高高在上遥不可攀。现代人力资源管理领域应在实际工作中充分利用这些技术和工具，创新管理手段，提升管理水平。所以，可以说数据分析是人力资源管理发展的趋势。

1.1.2　数据分析体现人力资源从业人员的技术刚性

老梁：经理，要达到您说的更高层次的数据分析水平，可能需要学习很多计算机和统计学知识，我担心这会阻碍人力资源管理人员去应用数据分析技术。

Miss 陈：对人力资源管理人员来说，要额外学习计算机和统计学知识确实有难度，但对于这些知识其实只需要学习基础内容就可以了，而基础内容的难度并不大。比如学习 R 语言，只需要掌握语法和数据结构等基础知识，就可以开始应用了。R 基本上是采用函数编程，很多算法模型往往就是那几个函数，设置一下参数就可以建模。用了之后你会发现和 Excel 的函数用法差不多，上手应该会比较快。统计学方面的学习也不用去研究算法原理，可以把算法当作黑匣子，只需要学习算法的输入、输出和适用条件等基础内容就足够了，这样其实比较简单。

老梁：学习基础知识恐怕也要花不少时间呢！

Miss 陈：学习当然需要付出时间和精力，不过一旦迈入数据分析的世界，你会发现人力资源管理迈上了一个新的层次，人力资源的管理水平和技术水平将显著提高，人力资源管理人员的技术刚性也将显著提高。到时你就会明白这种付出是非常值得的。

老梁：经理，您说的技术刚性是什么意思？

Miss 陈：刚性本来指物理属性，是物体承受外来压力但性质不发生改变的属性。这里说的技术刚性，指技术能力达到一定高度而不受外部变化影响的能力，也就是说技术能力达到了某种境界而表现出不可替代性。

老梁：明白了，您的意思是数据分析能够提高人力资源管理人员的技术能力，提高人力资源管理岗位的不可替代性。

Miss 陈：是的。你在公司时间也不短了吧，应该看到这几年常有人

员调到人力资源管理岗位工作，这些人员的专业出身五花八门，市场、财务、经营管理、综合、技术的都有，给人的感觉是什么人都能搞人力资源管理工作，这是什么原因造成的呢?

老梁：咱人力资源管理的工作给别人的感觉是技术门槛低，谁都可以来做。这和财务工作的对比最明显，不懂财务知识根本没法开展工作，但不懂人力资源管理知识也可以开展工作。

Miss 陈：这就是人们对人力资源管理的刻板印象，认为人力资源管理专业门槛低，入门容易，人人都可以做。但实际上我们都知道，人力资源管理涉及的知识范围非常广，能力要求也非常高。你看咱部门的本科、研究生占比，是全公司所有部门中最高的，这在某种程度上也说明了人力资源管理对人的能力要求很高。

要改变人们的刻板印象是相当难的，数据分析恰好可以成为改变印象的重要元素。这是因为数据分析代表了较高的知识和技术含量，具备技术刚性，一旦将人力资源管理与数据分析技术结合起来，某种程度上也提高了人力资源管理本身的技术刚性。

老梁：嗯，明白了，看来学习数据分析对人力资源管理人员来说是非常必要的。

1.1.3 数据分析能够为人力资源管理者提供强有力的决策支持

Miss 陈：当然，人力资源的数据分析最重要的作用还是给企业管理层提供决策依据。

老梁：就是说将分析结果提供给公司领导去做决策吗?

Miss 陈：是的，这点非常重要。如果数据分析只用于人力资源管理本身，只用于提高人力资源管理的水平，则显得狭隘了。若数据分析能给管理层提供有用的信息，能够影响和帮助公司做出正确的经营决策，才真

正体现了数据分析的价值。

比如,我们分析各个分公司的人力资源管理效能,分析分公司在人力资源管理投入和产出上的差异,再结合行业对标数据,对下一年的人员配置、工资分配提出相应的优化方案,将分析和方案提供给管理层,那么管理层就可以根据这些信息决定是否调整公司的经营指标和预算,更合理地给分公司下达经营任务等。这其中数据分析的内容就成为了重要的决策依据。

老梁:嗯,如果能引起管理层的重视,能够给管理层提供有效的信息,那也不枉咱们花时间去学习这些知识啊。

1.1.4　数据分析是人力资源管理的刚性需求

老梁:其实咱们天天都在接触数据,基本上各种总结、报告都会用到数据分析,虽然目前数据分析的层次还有待提高,但感觉数据分析已经是工作的一部分了。

Miss 陈:的确是这样,实际上我们的工作根本离不开数据。人力资源管理六大模块中,人力资源规划、招聘与配置、培训与开发、绩效管理、薪酬福利管理等模块都要以数据为基础,这些模块每天都会产生大量数据,加上各种管理系统及其存储的数据,可以说人力资源管理人员就是围绕数据在干活。

老梁:是啊,我们跟您汇报工作时如果没有数据来支撑内容,都不好意思拿出手,没有数据分析的报告也没有多少说服力。您看每个季度公司的经营分析会,都有人力资源分析,其中包含大量的数据分析,如人工成本、工资总额、人员流动情况等,都需要用数据来说话。

Miss 陈:所以进行数据分析并且不断提升数据分析水平是人力资源管理的刚性需求,是我们必须要做的工作。

1.2 人力资源数据分析有什么特点

1.2.1 数据分散性

Miss 陈：不过咱们人力资源管理用到的数据，可不是轻易就能得到的。

老梁：啊?! 咱们的数据不都是现成的吗，您看像薪酬、培训、绩效这些数据都在人力资源管理系统中，要什么数据都可以导出来，应该说还是比较容易得到的吧。

Miss 陈：这些数据自然可以轻松得到，因为这是我们的业务数据，但是进行人力资源的数据分析需要的不只是这些数据。比如，我们要做人力资源效能分析，就需要公司经营方面的数据，才能计算劳动生产率、人工成本创利、人工成本创收等指标；如果要做薪酬公平性分析，就需要了解外部行业薪酬数据；如果要进行人员流动性分析，就需要知道行业或岗位流动率对标数据。这些数据可不是那么轻松就能得到的，因为它们分散在各个地方。

老梁：噢，这么说来的确是这样。经营数据要到财务部、市场部去收集，外部数据要在网络上搜索，或者向咨询公司购买。这么说来人力资源分析所需要的数据是挺分散的。

Miss 陈：不仅如此，即便是在咱们部门内部，数据也是分散的。例如，招聘时应聘者的素质测评分数得找小肖，人工成本、工资总额、工资使用进度等数据得找小姚，培训记录、绩效考核的数据得找小曾。虽然咱们有人力资源管理系统，但培训、招聘等系统是独立的，薪酬数据由于需要保密也只能由专人管理，所以我们部门内部的数据也是分散的。

老梁：是啊，每次做经验分析我都得找小肖、小姚、小曾拿数据，要花不少时间才能集齐数据。

Miss 陈：人力资源数据分析的特点之一就是数据分散性。我们需要的数据都分散在相关人员、相关部门或者外部网络、机构中，在分析时需要花不少力气来收集、整理。特别是经营数据，涉及市场、财务等部门，这些部门可能会出于某些原因拒绝提供数据，所以数据收集的难度不小，即使收集了也不一定能获得理想的效果，给我们进行数据分析带来了一定的难度。

1.2.2　数据相关性

Miss 陈：人力资源数据分析的另一个特点是数据相关性。

老梁：相关性是不是指数据之间的关联性呢？

Miss 陈：是的，这种相关性体现在业务数据内部相关、与经营数据相关、与外部数据相关等方面。

比如，人力资源的业务数据中，培训、薪酬、绩效数据都是基于员工关联的，是员工产生的数据，彼此是相互联系的。

人力资源数据也受到经营数据的影响，比如公司经营效益好时，员工薪酬会上升，培训费用会增加，可能会多招聘员工；而经营效益不好时，则员工薪酬、培训费用下降的可能性较大，还可能会裁员，这说明人力资源数据和经营数据是也是相关性的。

老梁：明白了，经理，我来说说外部数据的相关性吧。我想到一点，我们的薪酬水平、人工成本等数据和政府发布的社平工资、最低工资、工资指导线等外部数据是相关的，比如社平工资上升，那么员工的社保、公积金的基数就会调整，会直接影响到公司的人工成本，这点就体现了人力资源数据与外部数据之间的相关性。

Miss 陈：说得很好。

1.2.3 非标准化数据

Miss 陈：人力资源数据分析还有个特点，这个特点会让我们特别头疼。

老梁：是什么特点呢？

Miss 陈：人力资源数据缺乏统一表征，从统计指标、统计口径到计算公式都缺少统一的标准。这个特点和财务数据形成了鲜明对比。财务数据标准化程度相当高，比如常见的资产负债表、利润表、现金流量表这三张报表的统计指标、口径、计算公式都是有统一标准的，每家企业都按照相同标准来计算和分析。对比起来，人力资源的数据就显得寒碜了不少。

老梁：咱们的劳动生产率、人均创利、百元人工成本创利、百元人工成本创收等指标都是标准口径的数据啊。

Miss 陈：不然。说起来人力资源统计指标挺多的，除了你说的这些，还有人工成本投入产出比、企业劳动分配率、人事费用率等，算下来也有百十来个指标，涉及人力资源的各个模块。但是这些指标并没有形成统一标准，其统计口径、计算方式在不同的企业或多或少有些差异。

首先是统计指标没有标准。比如，分析人工成本投入和产出，既可以用百元人工成本创利、百元人工成本创收，也可以用劳动分配律、人事费用率、人工成本占总成本费用比等指标，具体用哪些指标需要企业自己选择，所以不同企业可能有不同算法。

其次是统计口径没有标准。比如，最常见的劳动生产率，有些企业的统计口径是以与公司签订了劳动合同的员工来计算，有些企业则会将派遣员工合并计算，还有些企业可能会将外包业务的员工也统计进来。

老梁：咱们人力资源的数据确实存在这种问题，统计指标倒是多，但选用哪些指标，用什么口径来统计，每个企业的做法可能都不同，这的确

是一个让人头疼的问题。

1.3 大数据和人力资源管理的关系

1.3.1 人力资源数据是大数据吗

老梁：经理，现在不是已经进入大数据时代了吗，那么人力资源的数据分析属于大数据吗，能应用大数据的分析方法吗？

Miss 陈：人力资源的数据还算不上大数据，至少在咱们公司还没达到这个量级。大数据的特点是数据量大，达到 TB 甚至 PB 级别。1TB 的理论值等于 1 024GB，你想想咱们公司的人力资源数据有这么大的体量吗？大数据要用专门的工具来管理和分析，比如用 Hadoop（分布式系统架构）来管理，而我们的数据更多是用 Excel 来管理，从这点上看我们公司的人力资源数据也不是大数据。

老梁：哦，看来咱们没跟上大数据的趋势啊！

Miss 陈：虽然咱们的数据量级算不上大数据，但也可以跟上大数据的步伐，咱们做不到形似，但可以做到神似。

1.3.2 大数据技术可以用在人力资源管理上吗

老梁：您不是说咱们的数据算不上大数据吗，那怎么能做到神似呢？

Miss 陈：这和大数据的特点有关系，我们先来看看大数据的特点吧。大数据包括五个基本方面的内容。

（1）数据挖掘算法：大数据分析的理论核心就是数据挖掘算法，各种数据挖掘的算法基于不同的数据类型和格式才能更加科学地呈现出数据

本身具备的特点，也正是因为使用这些被全世界统计学家所公认的各种统计方法，才能深入数据内部，挖掘出数据的价值；也正是因为有这些数据挖掘的算法，才能更快速地处理大数据。如果一个算法要花上好几年才能得出结论，那大数据的价值也就无从说起了。

(2) 预测分析能力：大数据分析最重要的应用领域之一就是预测性分析，从大数据中挖掘出数据的特点，建立科学的模型，之后便可以通过模型带入新的数据，从而对可能发生的事情进行预测。

(3) 可视化分析：大数据分析的使用者有大数据分析专家，同时还有普通用户，但是他们二者对于大数据分析最基本的要求就是可视化分析，因为可视化分析能够直观地呈现大数据的特点，同时能够非常容易被读者所接受，就如同看图说话一样简单明了。

(4) 数据质量和数据管理：大数据分析离不开数据质量和数据管理，高质量的数据和有效的数据管理，无论是在学术研究还是在商业应用领域，都能够保证分析结果的真实性和有价值。

(5) 语义引擎：大数据分析广泛应用于网络数据挖掘，可从用户的搜索关键词、标签关键词或其他输入语义，分析、判断用户需求，从而实现更好的用户体验和广告匹配。

明白了吗？数据挖掘算法、预测分析能力、可视化分析这三项其实是大数据的精髓，是反映数据价值的关键。通过数据挖掘、预测和呈现，才能充分发挥数据的价值。而这三项其实和数据的大小没有太大关系，即便是咱们公司的小数据，也可以进行数据挖掘、预测分析和可视化。

老梁：哦，这是用了大数据的思想。

Miss 陈：是的。咱们再从技术上看一下吧，大数据用到的技术包括以下几个方面。

(1) 数据采集：将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。

(2) 数据存取：存取数据的工具包括关系数据库、NOSQL(泛指非关系型数据库)等。

(3) 基础架构：云存储、分布式文件存储等。

(4) 数据处理：通过自然语言处理让计算机“理解”自然语言。

(5) 统计分析：假设检验、显著性检验、差异分析、相关分析、T 检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、岭回归、logistic 回归分析、曲线估计、因子分析、聚类分析、主成分分析、因子分析、快速聚类法与聚类法、判别分析、对应分析、多元对应分析(最优尺度分析)、Bootstrap 技术，等等。

(6) 数据挖掘：分类、估计、预测、相关性分组或关联规则、聚类、描述和可视化、复杂数据类型挖掘(Text，Web ，图形图像，视频，音频等)。

(7) 模型预测：预测模型、机器学习、建模仿真。

(8) 结果呈现：云计算、标签云、关系图等。

以上大数据所用到的技术中，数据处理、统计分析、数据挖掘、模型预测、结果呈现都可以用在小数据上，也就是说可以用于人力资源数据分析中。

老梁：这么看来，虽然大数据的特点是数据量巨大，但是数据处理、统计分析、数据挖掘、模型预测、结果呈现等技术并不是大数据专用。明白了，咱们的确可以借鉴大数据的思想和技术，用于人力资源的数据分析，实际上还是赶上了大数据的潮流啊。

Miss 陈：是的。

1.4 人力资源数据分析的难点

1.4.1 取数难

Miss 陈：人力资源的数据分析还存在一些难点，这些难点会对我们的数据分析工作造成障碍。

老梁：是什么难点呢？

Miss 陈：首先是收集数据存在一定难度。之前说了人力资源数据具有分散性，这种分散性导致了收集数据存在困难。比如，我们进行人力资源效能分析的时候，需要收集公司的经营数据，包括合同量、工作量、收入、利润等数据，如果要做预测分析还需要历史经营数据，这需要向市场部和财务部取数，需要这两个部门的配合和支持，而且这些数据并不是现成的，需要花一些时间来统计，往往不能及时拿到，或不能拿到准确的数据。

老梁：还好，咱们公司的市场部和财务部挺配合咱们的工作，只要是出于工作原因，需要的数据基本都可以取到。当然有时候不能立即得到数据，因为有些数据他们也需要时间来统计，不过已经足够好了。

Miss 陈：是的，我们公司还好。不过听说有一些企业的经营数据可不是那么容易获取的，这和部门之间的沟通、协作程度有关系，协作程度不高的部门取数是比较麻烦的事情。

再比如我们进行薪酬公平性分析时，需要取外部的薪酬数据来对标，而这类薪酬数据没有现成的，在互联网上也很难搜索到，即使搜索到了也不敢轻易使用，因为不能保证数据的真实性。所以，薪酬数据一般需要向咨询公司购买。比较麻烦的是不同咨询公司的薪酬数据也不

一定相同，这是由咨询公司薪酬调查的方法、取样范围和区域不同等因素造成的。所以对于咨询公司出卖的薪酬数据，我们还需要明确数据的调查对象、调查范围和区域、调查方法等，以此才能决定是否能购买访该数据。

老梁：听上去的确比较麻烦。

Miss 陈：此外，获取人力资源的历史数据也有一定难度。人力资源管理往往重视数据的时效性，对当期数据比较敏感，很多分析是基于当期或同比数据，对更早的历史数据往往忽视，以致保存不周。在需要历史数据的时候难以短时间内获得，经常东拼西凑地寻找，花费了不少时间。

老梁：历史数据很重要吗？

Miss 陈：当然重要，数据挖掘中的很多算法都需要历史数据，比如回归分析，就需要大量的历史数据才能建立回归模型，进行分析和预测。

老梁：哦，真没意识到，看来咱们得定期整理历史数据，妥善保存，说不定哪天就能派上用场。

1.4.2　缺技能

Miss 陈：进行人力资源数据分析还有一个很大的障碍，就是人力资源管理人员本身的数据分析能力还不够高。

老梁：惭愧，俺也做了十多年人力资源管理工作，的确还不太会进行数据分析。不过也有客观原因，我在大学里没有学过数据分析，没有学过统计学，工作后也没有参加过相关培训，无从学起啊。

Miss 陈：是的，这不是你一个人的问题，大多数人力资源管理人员都存在这个问题，正是这些客观原因造成了人力资源管理人员中掌握数据分析技能的人很少。随着计算机技术的发展，统计技术和工具的普及，

以及大数据时代的到来，人力资源管理人员也要顺应当前发展趋势，主动学习和掌握一定的数据分析知识和技能，并将其应用到人力资源管理的实践中来，创造出人力资源管理领域的新天地，提升人力资源管理的水平，帮助企业更好地运作，实现经营目标。

老梁：经理，我和同事们一定会加强数据分析知识、工具的学习，提升我们的数据分析水平，提升我们的人力资源管理水平。

Miss 陈：好的，我们一起努力吧！

第 2 章

数据分析前的准备工作

导语：工欲善其事，必先利其器，选择合适的分析工具将让数据分析工作事半功倍。有了工具，还需要有材料，如何收集和清洗数据就显得至关重要，这也是整个数据分析过程中最消耗时间的工作。本章介绍各种数据分析工具，并通过对比分析重点介绍 R 语言这个数据分析的利器；然后介绍数据收集的工具和数据清洗的知识，这些都是进行数据分析前的准备工作。

2.1 如何选择数据分析工具

2.1.1 常用的数据分析软件

老梁：经理，俗话说“工欲善其事，必先利其器”。我们人力资源管理人员该如何选择一款合适的数据分析软件呢？

Miss 陈：数据分析的软件有很多，最常见的是我们熟悉的 Excel，除此之外还有许多专业的统计软件，带数据统计模块的计算机编程语言，带数据分析函数的数据库，等等。这些工具在其相关领域或行业中的知名度都很高，被广泛地应用在科研、商业等环境，比较著名且常见的数据分析软件有 R、SPSS、SAS、Matlab、Mathematica、Stata、Python、Eviews 等，如图 2-1 所示。

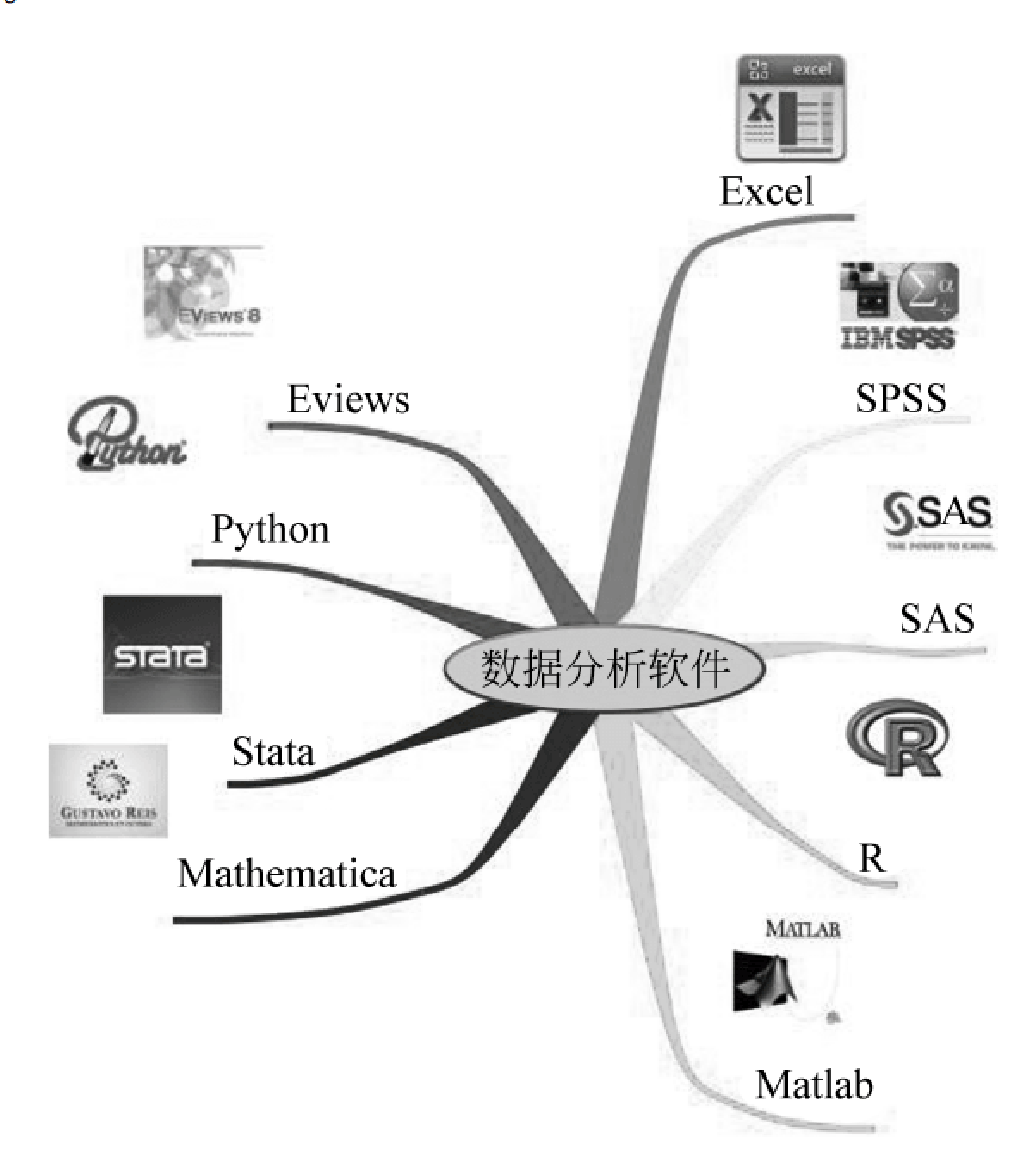


图 2-1 常见的数据分析软件

这些都是国内比较常见的数据分析软件。除了这些，其实还有很多数据分析软件，根据最新统计，数据分析软件有 93 款之多，涉及大数据、数据库、图表等方面，咱这里就不一一列举了。

老梁：您提到的这些软件，有些我听说过，比如 SPSS、SAS，但很多都没听说过。经理，这些数据分析的软件有什么特点呢？

Miss 陈：简单介绍一下刚刚提到的这些数据分析软件的特点吧。

(1) R：全称是 R language，即 R 语言。这是一种计算机语言，是专门用于统计分析、绘图的语言和操作环境。R 是一个免费、源代码开放的、跨平台的软件，是一个用于统计计算和统计制图的优秀工具。其功能包括数据存储和处理系统、数组运算(其向量、矩阵运算方面的功能尤其强大)、完整连贯的统计分析、优秀的统计制图功能、简便而强大的编程语言(可操纵数据的输入和输出)，可实现分支、循环，用户可自定义功能。从某种角度来说，R 语言的统计功能是所有统计软件中最强大的，因为除了传统的统计算法之外，目前最新的统计算法和研究技术都能在 R 语言中找到相关的函数包，几乎涵盖了人们在统计学领域的所有知识成果，而且算法更新速度极快，这点让商业领域的明星软件 SAS 和 SPSS 都望尘莫及。

(2) Excel：Microsoft Office System 中的电子表格程序，是我们经常使用的办公软件之一，使用频率非常高。它可以完成表格输入、统计、分析等工作，可生成精美直观的表格、图表，是我们日常工作中处理各种表格的首选工具。随着 Excel 的升级，现在还可以使用它跟踪数据，生成数据分析模型，编写公式以对数据进行计算，以多种方式透视数据，并以各种具有专业外观的图表来显示数据。由于 Excel 也有统计模块，所以可以说 Excel 也是数据分析软件。

(3) SPSS：全称 Statistical Product and Service Solutions，即“统计产品与服务解决方案”，IBM 公司的统计软件，可用于统计学分析运算、数

据挖掘、预测分析和决策支持任务的软件产品及相关服务。IBM 还有基于 SPSS 的衍生软件 SPSS Modeler，专门用于数据挖掘领域，提供了不少主流的数据挖掘算法（包括文本分析、实体分析、决策管理与优化）。SPSS 在生物、医疗、心理学等科研领域用得较多。

（4）SAS：全称 Statistical Analysis System，即“统计分析系统”，是由美国 NORTH CAROLINA 州立大学于 1966 年开发的统计分析软件，总部位于美国北卡罗来纳州的凯瑞，是全球最大的私有软件公司。SAS 系统在国际上已被誉为统计分析的标准软件，是全球商业智能和分析软件与服务领袖，全球 50 000 多家企业都在通过 SAS 软件对数据进行深入挖掘，在各个领域得到广泛应用。另外，SAS 可能是最贵的统计软件。

（5）Matlab：Matrix laboratory 的缩写，是一款由美国 The MathWorks 公司出品的商业数学软件，是一种用于算法开发、数据可视化、数据分析，以及数值计算的高级技术计算语言和交互式环境。Matlab 还可以用来创建用户界面及与调用其他语言（包括 C、C++ 和 Fortran）编写的程序。Matlab 主要用于数值运算，但利用为数众多的附加工具箱（Toolbox）它也适合不同领域的应用，例如控制系统设计与分析、图像处理、信号处理与通信、金融建模和分析等。另外还有一个配套软件包 Simulink，提供了一个可视化开发环境，常用于系统模拟、动态/嵌入式系统开发等方面。数学专业的同学们基本上都会学习这个软件。

（6）Mathematica：由美国科学家斯蒂芬·沃尔夫勒姆领导的沃尔夫勒姆研究公司（位于美国伊利诺伊州香槟市）开发的一款被广泛使用的计算软件。它拥有强大的数值计算和符号运算能力，是目前为止使用最广泛的数学软件之一。软件名字“Mathematica”还是由苹果创办人乔布斯向沃尔夫勒姆公司创立者提议命名的。Mathematica 和 Matlab 都是数学领域的主流软件。

(7) Stata：数据分析、数据管理以及绘制专业图表的整合性统计软件。Stata 的统计功能很强，除了传统的统计分析方法外，还收集了近 20 年发展起来的新方法，如 Cox 比例风险回归，指数与 Weibull 回归，多类结果与有序结果的 logistic 回归，Poisson 回归，负二项回归及广义的负二项回归，随机效应模型等。

(8) Python：一种面向对象、解释型的计算机程序设计语言，与 C++、Pascal 等计算机编程语言类似。它的主要特点是语法简洁而清晰、具有丰富和强大的类库、免费且开源、代码可移植性强，能够把用其他语言制作的各种模块（尤其是 C/C++）很轻松地联结在一起。Python 有专门的数据分析库，比如数据分析三件套 Matplotlib、Nunpy、Scipy，可以进行科学运算、数据分析和统计绘图。

(9) Eviews：Econometrics Views 的缩写，通常称为计量经济学软件包。软件本意是对社会经济关系与经济活动的数量规律，采用计量经济学方法与技术进行“观察”，也是专门从事数据分析、回归分析和预测的工具。使用 Eviews 可以迅速地从数据中寻找出统计关系，并用得到的关系去预测数据的未来值，其应用范围包括科学实验数据分析与评估、金融分析、宏观经济预测、仿真、销售预测和成本分析等。

2.1.2　选择数据分析工具的策略

老梁：经理，这么多数据分析软件让我眼花缭乱啊，好像个个都不错呢，该如何选择呢？

Miss 陈：不同的使用者应该考虑不同的选择策略，根据实际需求来选择合适的数据分析工具。我们是人力资源管理从业人员，那么就先分析一下我们在数据分析方面的需求和特点吧。

(1) 人力资源需要分析的数据量级不大，远未达到大数据量级。大

数据是指数据的体量很大，达到或超过 1TB 规模的数据，显然人力资源的数据没有达到这个级别，只是小数据。

(2) 人力资源需要分析的数据种类较多，涉及人力资源管理的各个模块。比如招聘、培训、绩效、薪酬等管理模块都会产生数据。由于我们全面启用了人力资源管理系统，这些数据多数都存储在数据库中，格式比较规范，并且容易收集。

(3) 人力资源的数据统计方法相对比较基础和传统，一般用计数、汇总、百分比、平均数等方法，从不同维度进行统计，通过同比、环比、横向对比、对标等方式进行分析。

(4) 人力资源管理的从业人员在数据分析方面所知所学不多，很多人在工作后才学习使用各种软件并接触数据分析。

老梁：经理，您说得对啊。

Miss 陈：所以，作为人力资源管理从业人员在选择数据分析软件的时候，应该根据我们的需求特点，从功能性、易用性、经济性三个维度，去衡量如何选择合适的数据分析软件。

老梁：您是说应该选择功能强大、简单易学并且成本又不会太高的分析软件？

Miss 陈：是的，其实就是选择性价比。在 Excel、R、SPSS、SAS、Matlab、Mathematica、Stata、Python、Eviews 等软件中选择的话，那么 Excel 在易用性方面比较突出，也有一定的统计分析功能，可以作为初、中级用户的选择；R 在功能性、经济性方面比较突出，可以作为中、高级用户的选择，如图 2-2 所示。

老梁：经理，您前面提到 Python 也是免费的，从功能性、易用性、经济性三个维度来看也有优势，为什么不选它呢？

Miss 陈：Python 虽然简单、强大、标准、免费，但它是一门计算机编程语言，它能做的事情太多，而数据分析只是它众多功能模块中的一个小

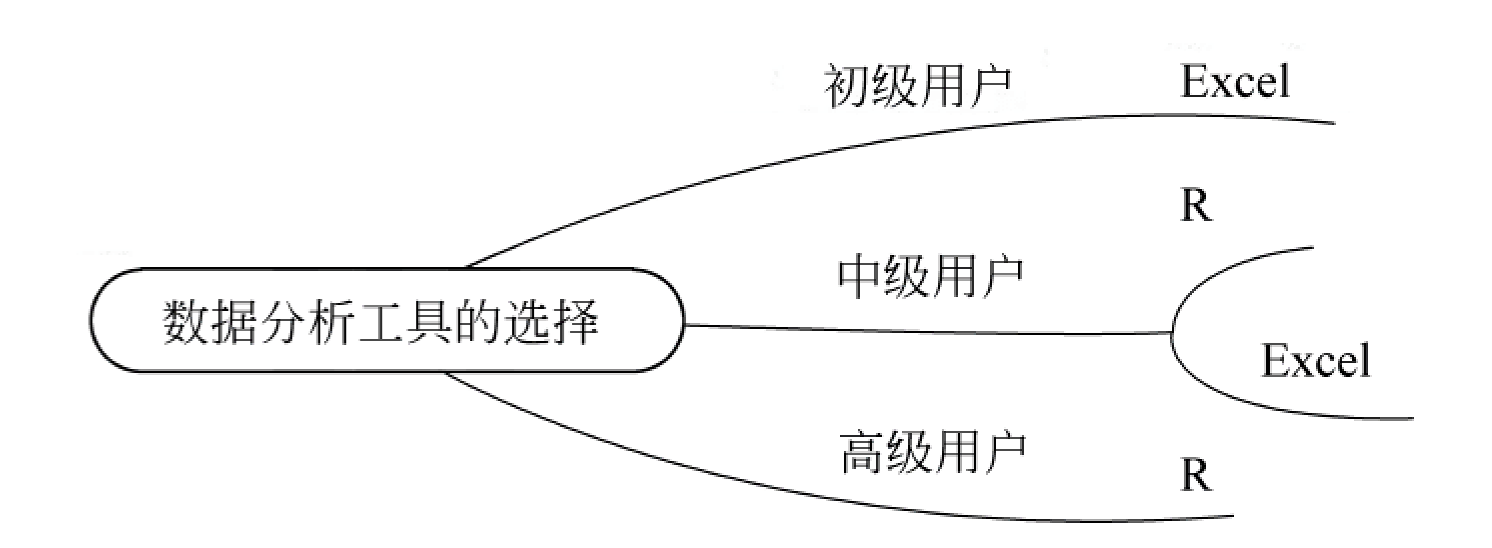


图 2-2　不同级别用户数据分析工具的选择

模块，不是其专长，我认为 Python 更适合计算机编程专业人士使用。相对而言，R 虽然也是一门编程语言，但 R 是专门用于数据分析的语言，其所有的功能都为数据分析而设计。所谓术业有专攻，在数据分析领域，R 语言更具优势，更适合我们去使用。

2.1.3　关于 Excel

老梁：经理，既然 Excel 在功能性和易用性上有优势，那么我们是不是用 Excel 进行数据分析就可以了啊？毕竟我们对 Excel 的熟悉程度高，上手容易，学习成本也较低。

Miss 陈：很遗憾，Excel 不能完全满足我们的分析需求。不过既然提到 Excel，那么我们就谈一谈它，因为对绝大多数人力资源管理人员来说，Excel 几乎是数据统计分析的唯一选择，日常工作中的数据分析基本都靠 Excel 来完成。

老梁：是啊，我们每天都在用 Excel 进行数据统计和报表制作。

Miss 陈：所以 Excel 是我们最常使用的办公软件之一，使用频率非常高，甚至可以说是 office 办公软件中使用频率最高的软件。而且不仅是咱们人力资源部，公司的各个部门都会用到它，比如市场部做经营分析、财务部做财务分析等，都会使用 Excel。

从功能上来讲，Excel 可完成表格输入、统计、分析等工作，可生成精

美直观的表格、图表，是我们日常工作中处理各式各样表格的优秀工具。并且随着 Excel 的升级，新的版本还可以跟踪数据，生成数据分析模型，编写公式以对数据进行计算，以多种方式透视数据，以各种具有专业外观的图表来显示数据，数据还可以存储到云中保存。

在数据分析方面，Excel 提供了一套分析工具库和用于数据分析的 VBA 函数库，可以比较方便地进行一些高级的统计分析，比如常见的回归分析、t 检验、F 检验、方差分析、计算相关系数等，都可以在 Excel 的数据分析库中找到，如图 2-3 所示。但是 Excel 提供的这些数据功能相比专业的统计分析软件来说，具有种类不多、计算结果简单、图形粗糙等缺点，不过据说用 Excel 提供的 VBA 函数也能实现很多数据分析算法，但需要编写大量代码，会很耗时间。

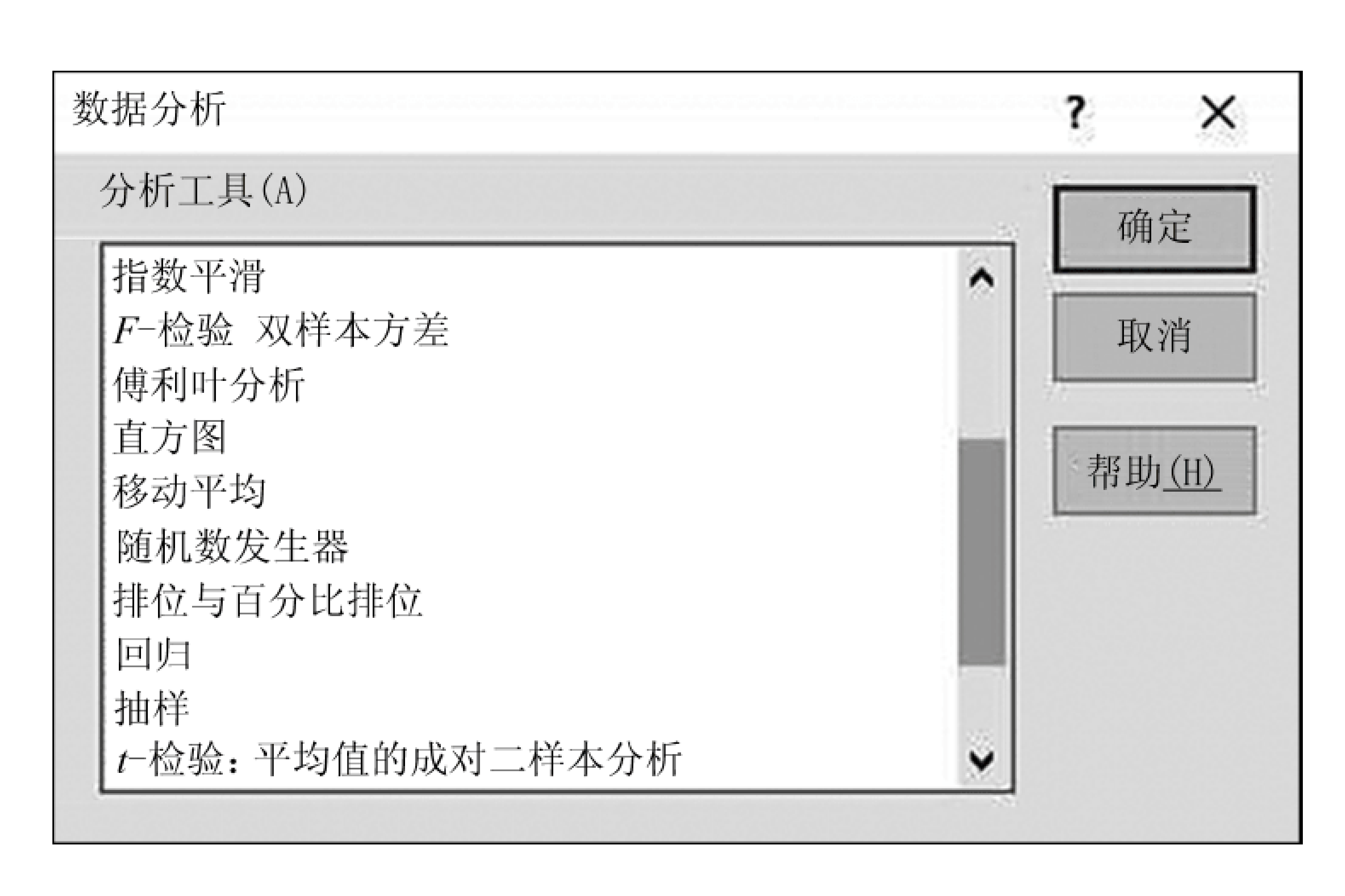


图 2-3 Excel 中的数据分析工具

老梁：经理，虽然如此，但关键是别的软件咱也不会啊，Excel 的功能如此强大，又容易上手，所以自然就想到 Excel 了。您说 Excel 还可以进行回归分析之类的统计分析，感觉很不错呢。

Miss 陈：是的，Excel 老少咸宜。打个比方，Excel 就像一把菜刀，人人都可以用来切菜，但是不同的人有不同的用法，会产生不同的效果。普

通的人仅仅用来切菜，厉害的人可以用菜刀杀猪宰羊，各样操作游刃有余。类似的，Excel 用到高深之处，一切和数据相关的工作都可以胜任，甚至还可以用它来编写游戏。

老梁：看来我的 Excel 运用还处于初级阶段，只会用来进行简单的数据统计和做报表，从来没碰过数据分析工具、VBA 这些东西，哈哈。

Miss 陈：虽然 Excel 有许多优点，但也有不足的地方。

(1) Excel 的高级数据分析功能比较简单。虽然 Excel 提供了分析工具库，但功能却比较简单。比如回归分析，若要进一步进行自变量多重共线性的检验，就做不了，也不能做逻辑回归分析。Excel 提供的分析算法也不多，诸如分类、降维、非参数检验等算法都没有，更别说当前大数据时代流行的机器学习算法。虽然有 Excel 的 VBA 可以编写代码，但难度是非常大的。

(2) Excel 的绘图功能还不够强。大家可能都对 Excel 默认的图表功能抱怨过，特别是 2003 年版及以前的版本，实在缺乏美感。虽然新版本的 Excel 图表好看了很多，还加入了应用商店，可以绘制一些流行的图形，但我认为 Excel 的绘图还是不够强大，绘制复杂图形时需要进行烦琐的设置，并且绘制多张复杂图表时操作显得更加烦琐。Excel 的绘图功能跟它的统计功能类似，基本的功能都有，很容易就可以绘制基本图形，但是复杂图形就需要研究很久、设置很多参数。

(3) Excel 是微软 Office 办公套件的一部分，价格不菲。虽然家庭版、学生版比较便宜，但很多功能都被阉割了。比如 Power View 功能在家庭版和学生版上就找不到，必须得用专业版，可是专业版的价格就很高。现在微软的收费方式又有变化，采取月费或年费的方式，每年都得花钱，算起来开支不小。单就软件收费而说无可厚非，但成本费用一定是影响我们选择软件的重要因素。

老梁：是啊，咱们公司当年买 Office 办公软件可花了不少钱，不过随着时间推移，以前的版本都过时了，还没升级呢。看别人新版本的 Excel 界面很酷，功能很多，可惜咱们没得用啊。如果要用得花不少钱呢，大家都在等着公司升级 Office 版本，可是不知道啥时候才会升级到新版本。

Miss 陈：呵呵，公司如果升级 Office，那将会是一笔不菲的开支。现在版本的 Office 还能用，而且也不影响公司正常的经营生产，升级的必要性不大，所以公司多半会继续使用现在的版本。

2.1.4 关于 R 语言

1. R 语言的江湖地位

老梁：经理，俗话说，天下没有免费的午餐，像 R 语言这样免费的数据分析软件会不会有缺陷，如功能不全、性能不强，又或者有某些功能要收费呢？

Miss 陈：人们对免费的东西持有怀疑态度是一种常见的思维定式，就像超市里面免费品尝的东西实际上是在引诱你买货架上的产品，培训机构请你免费听课无非是进行广告宣传吸引你去参加收费的培训，旅游公司的免费旅游实际上会让你在购物点度过大部分时间。

但是在互联网领域、科学界，分享是一种价值观。在这种价值观的引导下诞生了一些高质量的免费软件，R 语言就是其中的佼佼者。R 语言是上帝给我们的珍贵礼物，你可以用 R 语言做一切数据统计分析方面的事情，尽情享用几百年来人类在数据统计方面的研究成果，各种算法应有尽有。最新的统计方法发表出来后通常会在 R 语言中率先实现应用，这让其他所有统计软件黯然失色。R 语言在数据分析、数据挖掘领域的功能之强大，胜过前面提到的任何一款统计软件，并且使用这一软件不需要花一分钱。

国外著名的数据分析和挖掘社区 KDnuggets 每年都会做一次关于数据分析、大数据、数据挖掘、数据科学使用软件工具的调查，根据 2015 年的调查结果，R 语言在参与评选的 93 款相关软件中排名第一，使用率达到了 46.9%，江湖老大的地位俨然确立。排名前 10 的数据分析软件如图 2-4 所示。

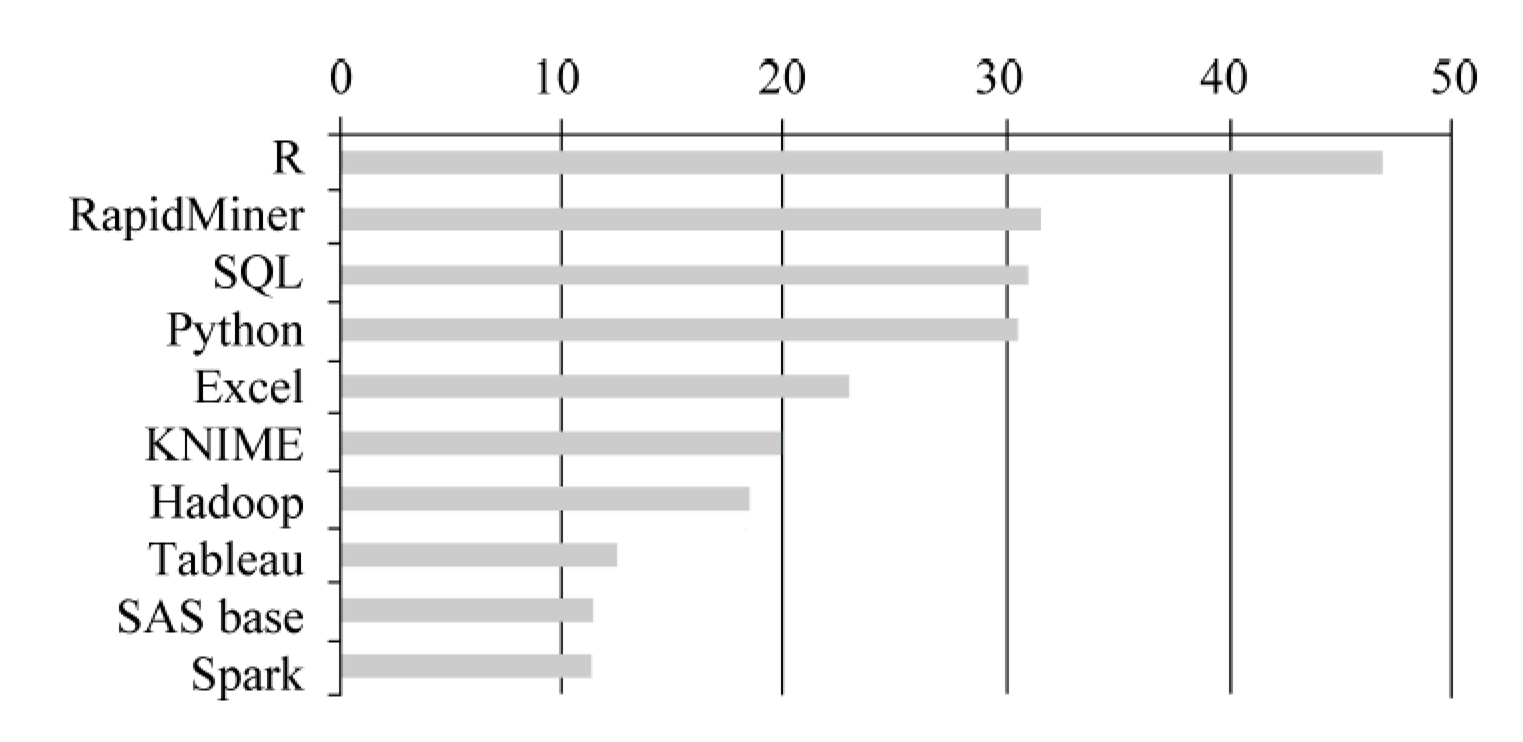


图 2-4 排名前 10 的数据分析软件(KDnuggets,2015)

老梁：真没想到，一个名称看上去如此简单、普通的软件，在数据分析领域的地位竟如此之高。

Miss 陈：其实 R 语言诞生得很早，之前一直在科研、专业领域传播和应用，随着大数据的流行才真正进入大众的视线。

2. R 语言的前世今生

老梁：经理，我很好奇 R 语言的来历。

Miss 陈：那给你讲讲 R 语言的故事吧。

R 语言源于 S 语言，S 语言也是一种用于统计分析的计算机语言。S 语言非常厉害，1998 年美国计算机协会(ACM)给 S 语言的设计者发了一个奖：软件系统奖，用来表彰 S 语言取得的成就。这个奖很牛，因为得奖的都是系统级别的软件，比如 Unix、TeX、TCP/IP、Word-Wide-Web、Java 等，个个大有来头。在所有获得软件系统奖的软件中，S 语言是唯一一个

统计软件，可见其厉害之处。不过S语言是商业软件，跟SPSS、SAS一样，需要花钱购买。

1993年，新西兰奥克兰大学的两位统计学家，一位叫Ross Ihaka，另一位叫Robert Gentleman。他们两位志趣相投、心意相通，利用业余时间对S语言进行了改进，创造出了一种新的统计语言。由于两位统计学家的名字都是以R开头，这个新的统计语言也就顺理成章被命名为R。

当年这两位大牛将刚诞生不久的R语言放到了卡耐基·梅隆大学的计算机服务器上，供大家下载研究。这时用R语言的人极少，但也有不少人进行了下载研究，其中来自苏黎世理工学院的一位学者在用了R语言之后，大力劝说两位作者公开源代码，让R语言成为自由软件。两年后，即1995年，两位教授本着分享、协作的精神，将R语言源代码正式发布到自由软件协会的FTP服务器上，自此R语言正式以自由软件的身份面向全世界。

随后的20年，R语言充分体现了互联网时代国际化协作发展的特点：诞生于新西兰，邮件列表维护在瑞士，服务器架设在奥地利，Windows版本主程序维护在加拿大，附加包维护在德国，Mac OS版本维护在美国，全球近20个国家有镜像网站。核心开发团队有20人，成员来自世界各地的大学，如牛津大学、加拿大西安大略大学等，也有来自企业的成员，比如AT&T实验室的Simon Urbanek等。

就是这样一种组织、维护形式松散的计算机语言，依靠着志愿者坚持不懈的贡献，在不断发展和升级。现在世界各地大量的优秀统计学家、各个领域的统计学爱好者、计算机程序员都在为R语言贡献自己的力量，将大量统计方法以附加包(package)的形式发布出来，使其他不擅长编程的用户能以最快的速度用上最新的统计方法。

2012年，R语言可以下载的package达到3 200个，用了17年；2015年，R语言可以下载的package翻倍达到6 800个，仅用了3年。那些封

闭源代码的商业统计软件很难有这样的发展速度，只能望尘莫及。R 语言像滚雪球一样，依靠开源、分享、协作的方式，从开始不温不火，蓄积能量，到后来逐渐显示出威力，再到大数据时代彻底爆发，成就了 R 语言的今天。

老梁：大开眼界了，没想到还有这样的软件，它就是由跨国界、跨种族的精英共同创造的智慧结晶啊，不仅免费，还集全世界各领域数据分析家的努力和智慧于一身，真是一个伟大的软件。听了 R 语言的故事，我已经被 R 语言深深吸引了，等下我就去下载 R 语言，马上安装，马上学习。

3. R 语言是算法聚宝盆

老梁：对了，经理，R 语言中的 package 都是用来做什么的？

Miss 陈：这些 package 是函数包，是为了解决某个问题或为实现某种统计算法而编写的函数集。在 package 中蕴藏着大量的统计算法，就像是一个聚宝盆，包含我们能想到的和不能想到的、学过的和没学过的、古老的和现代的、简单的和复杂的算法，应有尽有，可以称其为算法聚宝盆。其中部分算法如图 2-5 所示。



图 2-5　R 语言中的统计算法

老梁：经理，您说的算法是指什么呢？

Miss 陈：算法可以简单理解为解决问题的计算方法。例如，我们每个月要给员工发工资，按照《中华人民共和国个人所得税法》，要计算每个员工的个人所得税，而个人所得税实行累进税率，其计算公式如下：

应纳个人所得税税额＝应纳税所得额×适用税率－速算扣除数

上面的公式就是一个算法。如果编个函数，把这个公式用计算机语言来表示，再输入本月工资数额，计算出个人所得税，那么这个公式就可称为一个计算机算法。

老梁：哦，R 语言中的 package 就是这些算法的集合吗？

Miss 陈：是的。R 语言中的 package 包罗万象，包含了各种各样的算法，涉及数据分析的各个领域，比如生物、经济、金融、心理学、医学、人工智能，等等。

老梁：涉及的范围真广啊！

Miss 陈：现在流行的大数据分析，其背后也是各种统计分析算法在支撑，而不仅仅是简单的一些数据统计。比如你在浏览淘宝网页的时候，有没有注意网页的广告、推荐的商品，都符合你本人的购买倾向和喜好呢？

老梁：哎呀，是的，我正奇怪呢。最近上淘宝，看到有个页面叫“发现好货”，里面推荐的商品都是我最近浏览过的，或者是和我浏览过的商品相关的商品，更有我没有浏览过但觉得还不错的商品。所以打开这个页面后不由自主就看了好久，一不小心就买了不少东西。

Miss 陈：这是由于淘宝的大数据分析做得很好，后台有算法在分析用户的购买倾向。比如可以根据用户的注册资料，将用户的购买行为进行分类，用分类算法建立预测模型。当你注册淘宝用户时，会填写个人资料，这些资料包括你的性别、所在地、年龄、职业、学历等，对吧？你填写的资料越详细，淘宝对你的分析就越精准。淘宝可以分析这些资料，根据你购买商品的行为建立预测模型，就能预测你的购买倾向，计算出你购买不

同种类商品的概率是多少。当你再次浏览淘宝网页的时候,就会有针对性地向你推送你可能购买的商品,自然就能最大限度地激发你的购买欲望,购物成功率就会提高很多。

老梁:哎呀,原来是这样,看来以后资料不能填得太详细,否则自己的想法都被别人知道了。

Miss 陈:银行对用户申请信用卡、贷款、股票账户开户的风险评估基本也是用这类算法来实现的。这类算法有不少呢,在 R 语言中都能找到对应的 package,使用相当方便。很多商业软件才有的算法,比如神经网络、贝叶斯分类、决策树、随机森林、结构方程模型等,在 R 语言中都可信手拈来。

老梁:经理,R 语言中有没有咱们人力资源管理领域的 package 呢?

Miss 陈:人力资源属于管理领域,很少进行数据分析方面的研究,对算法的依赖性也不强,所以没有专门对应的 package。但是现今的人力资源管理领域亟须提升数据分析水平,以应对当前大数据技术发展的趋势,从而提高人力资源管理水平。为此,我们应该积极挖掘人力资源数据价值,尝试将数据分析的方法引入工作实践中,创新我们的管理方法,解决管理中出现的问题。实际工作中,我们可以根据具体问题具体分析,明确数据分析方面的解决方案,然后再去寻找对应的算法。

老梁:明白了,我们做人力资源管理的对这些算法知之甚少,看来以后得加强学习啊。

4. R 语言是绘图专家

老梁:您刚才提到 R 语言可以绘图,这方面 R 语言有什么特别之处吗?

Miss 陈:R 语言的绘图功能很强大。本来绘图只是 R 语言附带的功能,但得益于 R 语言的开放性,许多人又给 R 语言开发了专门的绘图

包，使得 R 语言的绘图功能变得异常强大，几乎不输于任何商业数据绘图软件。图 2-6 列出了部分利用 R 语言绘制的数据图，你可以看看。

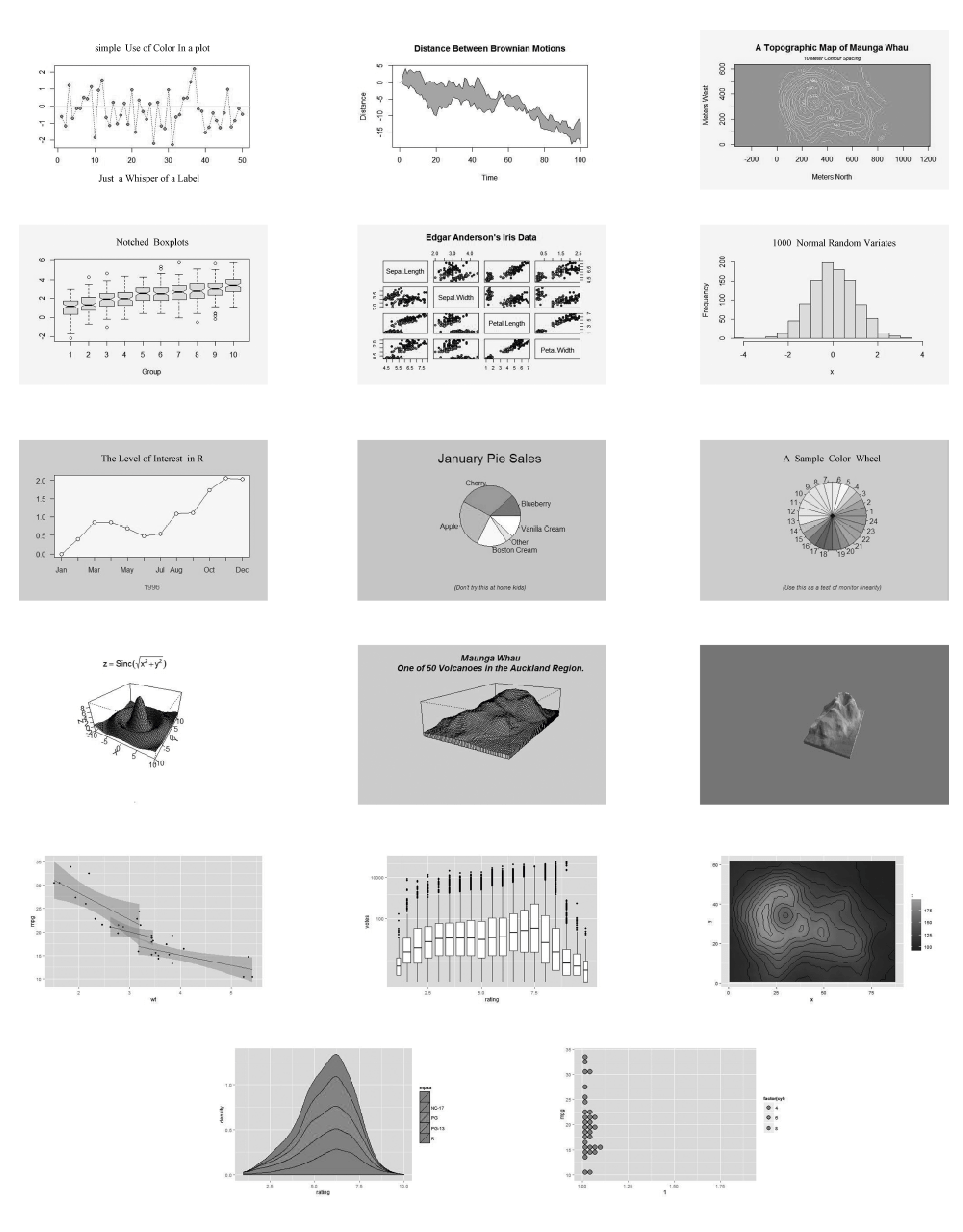


图 2-6　R 语言绘图功能展示

老梁：哇，这些图形看得我眼花缭乱，都是用 R 语言绘制的吗？

Miss 陈：是的，上面列出的图形只是 R 语言绘图功能的冰山一角，你可以上网搜索一下，能看到更多的 R 语言绘制的数据图形。其实 R 语言本身的绘图功能已经不弱，再加上许多人开发了功能更加强大的绘图包，提供给 R 语言用户使用，所以 R 语言的绘图功能变得强大。其中比较重要的绘图包有 ggplot2、lattice 等。ggplot2 包更是将 R 语言的绘图功能发扬光大，它将简单的图形语法融入 R 语言，使 R 语言能够绘制出各种惊艳、漂亮的统计图形，极大地扩大了 R 语言在图形领域的影响力。

5. 人力资源管理人员使用 R 语言的技能需求

老梁：经理，我觉得 R 语言的功能太强大了，package 浩瀚如海，作为人力资源管理人员，应该掌握 R 语言的哪些知识和技能呢？

Miss 陈：根据人力资源管理人员的特点，建议按照以下顺序学习 R 语言相关基础知识。

(1) 语法。其实 R 语言的语法很简单，多数时候几个函数就可以解决问题，并且这些函数用起来和 Excel 中的函数很相似。像循环控制、条件语句等都很少用到，除非要编写函数，但通常不需要这么做。R 语言不需要很长的代码，一个函数加几个参数就能制作一个复杂的统计模型，是比较典型的函数式语言。

(2) 数据类型和数据读取方法。R 语言中的数据类型有几种，最常用的是数据框(dataframe)，很多统计分析都是基于数据框来进行的。数据框的数据结构和数据库中的数据表类似，第一行是字段名，从第二行开始是记录，每个字段(每列)可以是不同类型的数据。然后，还需要掌握数据读取的方法，比如怎样从 Excel 中读取数据到 R 语言中。

(3) 绘图。R 语言绘图功能相当强大，一个 plot 函数就可以变化万千，绘制很多种图形。但强烈建议学习 ggplot2 绘图包，这种语法简单的

绘图方式，一旦接触使用就会被吸引，使你再也不想离开 R 语言。

掌握上述三方面的知识，就具备了用 R 语言进行数据分析的基础能力。这些内容在许多介绍 R 语言的书中都可以学到。具备这些基础后，根据实际工作需要，结合具体问题寻找对应的算法包，就能够进行数据建模、数据分析等操作了。

比如，在实际工作中，我们发现大学生入职后一年内的离职现象比较突出，给公司造成了不良的影响，增加了员工招聘的成本，于是想到能否在招聘前就预测出大学生在入职一年内的离职概率，从而提高招聘的质量。带着这个问题，我们就可以去寻找相应的分析算法，结果发现逻辑回归、决策树、Boosting、随机森林、神经网络等算法都可以实现这个目的。于是我们可以选择其中一种算法，下载对应的 package，学习其使用方法，研究其函数如何使用、对数据的要求、结果的解释，然后导入数据就可以进行分析和预测了。

老梁：那么是不是还要学习统计学方面的知识呢？

Miss 陈：当然，不过这方面知识的学习曲线会很长，涉及数学、概率等内容，对人力资源管理人员来说有不小的难度。比如，专门讲解贝叶斯分类的书就有好几本书，作为人力资源管理人员，学习这些算法原理几乎不可能，因为我们没有时间、精力和基础。合适的做法是，阅读一些统计学方面的科普书籍，了解常见算法的作用、适用条件、数据要求、结果解释等内容，也就是了解算法的基本原理、数据输入和结果输出。我们可以把统计算法当成一个黑匣子，就像我们看电视，仅需知道如何使用遥控器开关电视，如何选择频道即可，不需要去知道电视机内部的结构和实现原理。

老梁：这么说来我就松了口气，看来要学习 R 语言也不像想象中那样困难，掌握基础的内容后，有选择性地学习 package 的用法，就可以在实际工作中使用了。

Miss 陈：是的。对人力资源管理人员来说，诸如 Excel、Word 这样所见即所得的工具用惯了，要编写代码的确有障碍，其中最大的恐怕是心理障碍。不过，一旦克服了心理障碍，迈过这道坎，就会有令人振奋的收获，会发现新的世界。

2.2 如何有效收集数据

2.2.1 打通关节，从内外部渠道收集数据

Miss 陈：前面聊了数据分析的工具，接下来我们看看如何收集数据。我们要做数据分析，数据是最基础的东西，它是原材料，而原材料的获取非常重要，很大程度上决定了我们可以进行怎样的分析以及分析的质量。老梁，你说说平常我们的数据是从哪些渠道收集的？

老梁：我们的数据，一是从人力资源管理系统上获取；二是各单位上报；三是在网上搜索下载，大概就这三种方式吧。

Miss 陈：其实数据来源的渠道有很多，可以分为内部渠道和外部渠道，如图 2-7 所示。

2.2.2 内部渠道如何收集数据

Miss 陈：从图 2-7 中可以看到，收集数据的内部渠道主要包括以下方面。

(1) 企业内部各种信息化管理系统，这是最重要的数据来源。内部系统包括人力资源管理系统、财务管理系统、企业资源管理系统、OA(办公自动化)系统、项目管理系统等。有些系统还有若干相对独立的子系统，

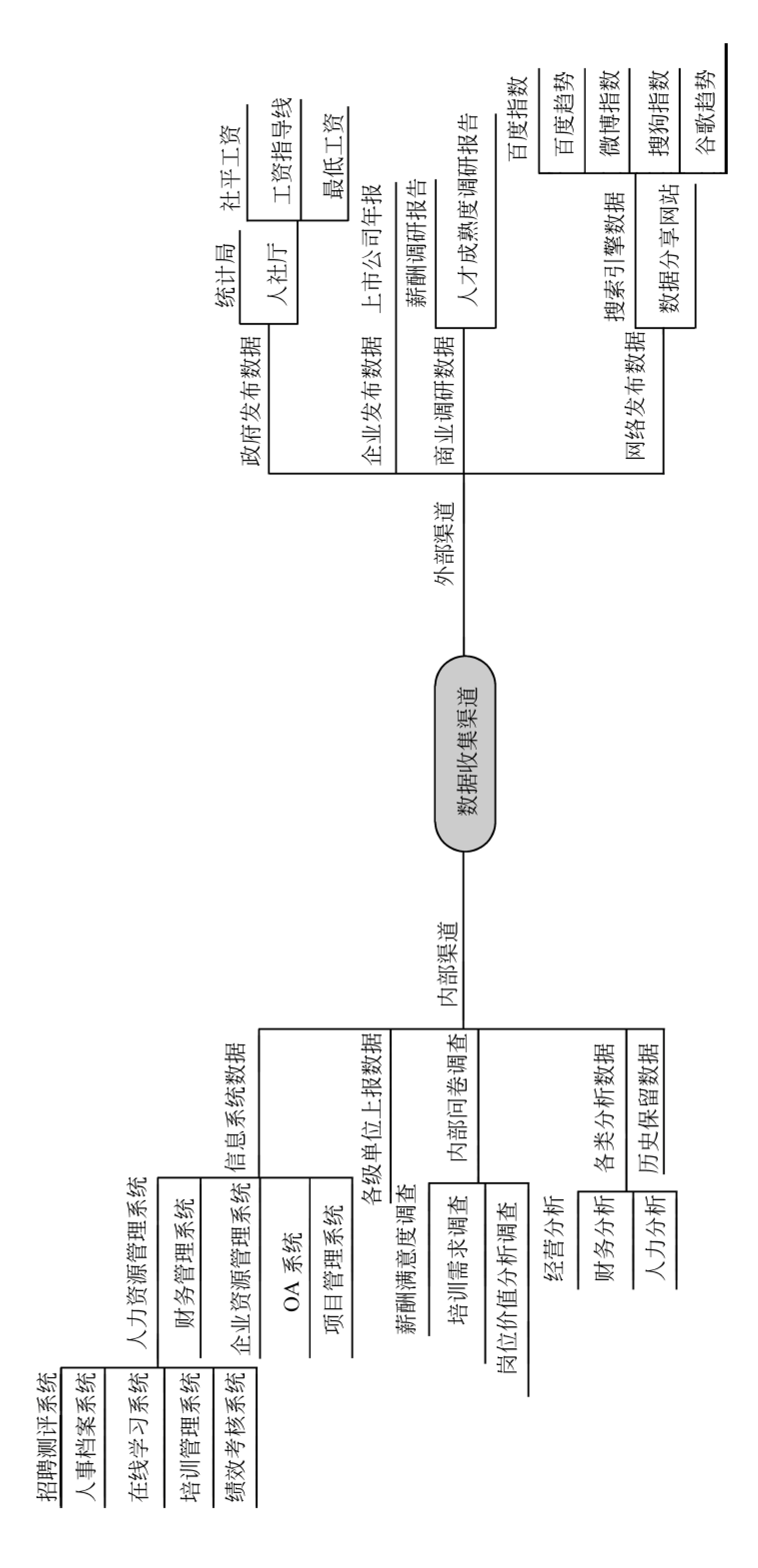

数据收集渠道
外部渠道
政府发布数据
统计局
人社厅
社平工资
工资指导线
最低工资
企业发布数据
上市公司年报
商业调研数据
薪酬调研报告
人才成熟度调研报告
网络发布数据
搜索引擎数据
百度指数
百度趋势
微博指数
搜狗指数
谷歌趋势
数据分享网站
内部渠道
信息系统数据
人力资源管理系统
招聘测评系统
人事档案系统
在线学习系统
培训管理系统
绩效考核系统
财务管理系统
企业资源管理系统
OA 系统
项目管理系统
各级单位上报数据
内部问卷调查
薪酬满意度调查
培训需求调查
岗位价值分析调查
各类分析数据
经营分析
财务分析
人力分析
历史保留数据

图 2-7 数据收集的渠道

比如我们的人力资源管理系统，下面还有招聘测评系统、培训管理系统、在线学习系统、人事档案系统和绩效考核系统。这些系统中存储了大量的数据，而且这些数据的存储形式都很规范，所以数据质量相当高。

(2) 各级单位上报数据。这是获取数据的简易渠道，在总公司层面，只需要发布通知，各单位就会按照要求填报数据上来，收集数据的速度较快。但这种方式的缺点是数据质量参差不齐，特别是数据格式、数据类型容易出错，造成后期数据清洗的时候会花不少时间。

(3) 内部问卷调查数据。比如，我们曾经做过的薪酬满意度调查、培训需求调查、岗位价值分析调查等，通过问卷的形式收集数据，也是一种简便有效的方式。

(4) 各类分析数据。主要是公司各个部门、分公司的分析报告，比如经营分析、财务分析和人力资源分析，这些分析切合企业实际，数据价值非常高。

(5) 历史保留数据。就是以前的数据，比如我们历年的招聘测评数据、人员流动数据、劳产率数据等，这些数据都非常有用，如果要做回归分析就需要积累大量的历史数据。

老梁：经理，我们从人力资源管理系统上收集数据就行了，为什么还要从财务、市场、项目管理系统去收集数据呢，还要研究它们的分析报告？

Miss 陈：人力资源数据一定要和市场、财务、项目管理等数据结合起来分析，才能体现价值，才能贴合企业经营发展的实际情况，才能更好地服务于公司的战略决策，否则就是闭门造车，没有说服力。

老梁：原来如此。

2.2.3　外部渠道如何收集数据

Miss 陈：再说说收集数据的外部渠道吧。

（1）政府发布数据。与人力资源相关的主要是统计局、人社部等官方网站的公开数据，政府会定期公布各种统计数据，比如每年的社平工资、工资指导线、最低工资等数据。但这些数据相对比较宏观，范围比较大，应用起来有一定难度。比如社平工资，实际上并不能反映大多数人的收入水平，因为是用平均数来代表收入水平，极容易受高收入者的影响而被拉高，而用中位数来代表薪酬则会更客观些。所以这类数据使用的时候需要慎重。

（2）企业发布数据。上市公司都会发布年报，其中一些数据是可以利用的，比如可以通过年报数据知道企业的劳动生产率、人均利润等，这类数据在各种财经网站都可以查到。

（3）商业调研数据。就是咨询公司通过调研收集整理的数据，这类数据最精准。比如，一些招聘网站或者咨询公司通过调研编制的年度薪酬报告，就非常符合企业的需要，可以直接用来对标，进行内外部薪酬的比较分析。我们公司就曾经购买过某大型人才网站的薪酬报告，用来优化某些岗位的薪酬水平。

（4）网络发布数据等。主要是各类网站、搜索引擎发布的数据，比如谷歌趋势、百度指数、百度趋势、微博指数、搜狗指数等，这些数据通常是免费的，登录相关网站即可查询。另外有一些行业网站也会发布很多数据，比如经济类、金融类网站会发布大量经济、金融、股票方面的数据，不过这类数据与我们人力资源管理工作的相关性不大。

老梁：外部的数据收集渠道的确很多，但是要获得薪酬方面的数据，好像必须购买商业调研数据才行啊。

Miss 陈：是的，薪酬数据的收集需要耗费较大的人力、物力和财力，成本较高。通常咨询公司都是将其作为产品售卖，价格不菲。不过最近几年互联网兴起了一些晒工资的网站和软件，借助网络的力量，许多人上传了他们的工资数据，供其他人查询和参考。这些数据的可信度值得商

权，但也可以作为参考数据。

2.3 与时俱进，运用各种工具收集数据

2.3.1 用 Adobe Acrebat 制作 PDF 问卷收集数据

Miss 陈：当我们需要通过问卷调查来收集数据的时候，通常我们会编制问卷，然后发给相关单位或人员，填写后回收整理，是吗？

老梁：是的，不过问卷调研比较麻烦，因为收集和整理数据要花很多时间。

Miss 陈：一般是怎么做的？

老梁：我会用 Word 先设计问卷，用 Excel 发布问卷和统计数据。相对来说 Excel 问卷的数据比较好收集和汇总。不过也挺花时间的，当回收的问卷达到几百份的时候，打开每个文件进行复制、粘贴的操作要花不少时间，是个体力活。

比如，图 2-8 所示的问卷，左边是 Word 版本，右边是 Excel 版本，就是我每次进行问卷调查做的样式。

Miss 陈：除了 Word 和 Excel，其实还有一种可以用来做问卷调查的文件类型，能够高效地进行问卷数据的收集。你知道 PDF 格式的文件类型吗？

老梁：知道啊，我们 OA 的公文都是 PDF 格式的。难道 PDF 格式的文件可以做成问卷吗？这种文件好像只能浏览，不能编辑啊！

Miss 陈：普通的 PDF 文件的确不能编辑，但是如果在里面添加一些可以编辑的元素，那就不一样了，就成了可以编辑的 PDF。还记得你去年申请澳大利亚签证时填写的申请表吗？看看是不是如图 2-9 所示。

Word 版本

**公司薪酬满意度调查问卷

调查问卷说明：

- 本调查问卷共有 48 个问题，问题采用单项选择的方式，简明扼要并易于回答。
- 你可以匿名填写此份调查表。
- 本调查问卷的保密级为 A 级，任何信息都将严格受到保密，所以你可以放心做答。
- 当有超过 50%的题目不做回答时，本问卷将做无效处理。
- 请你按实际情况作答，否则将影响调查结果。

你的姓名：________ 所在部门：________

你的职位：________ 入职年限：________ 你的年龄：________

性　别：________ 学历程度：________

1. 你对自己努力付出与工资回报二者公平性的感受是

（A）完全公平
（B）基本公平
（C）不确定
（D）不公平
（E）非常不公平

如果选择 D 或 E，请写明简要理由或感受：________________。

2. 以自己的资历，你对自己的工资收入

Excel 版本

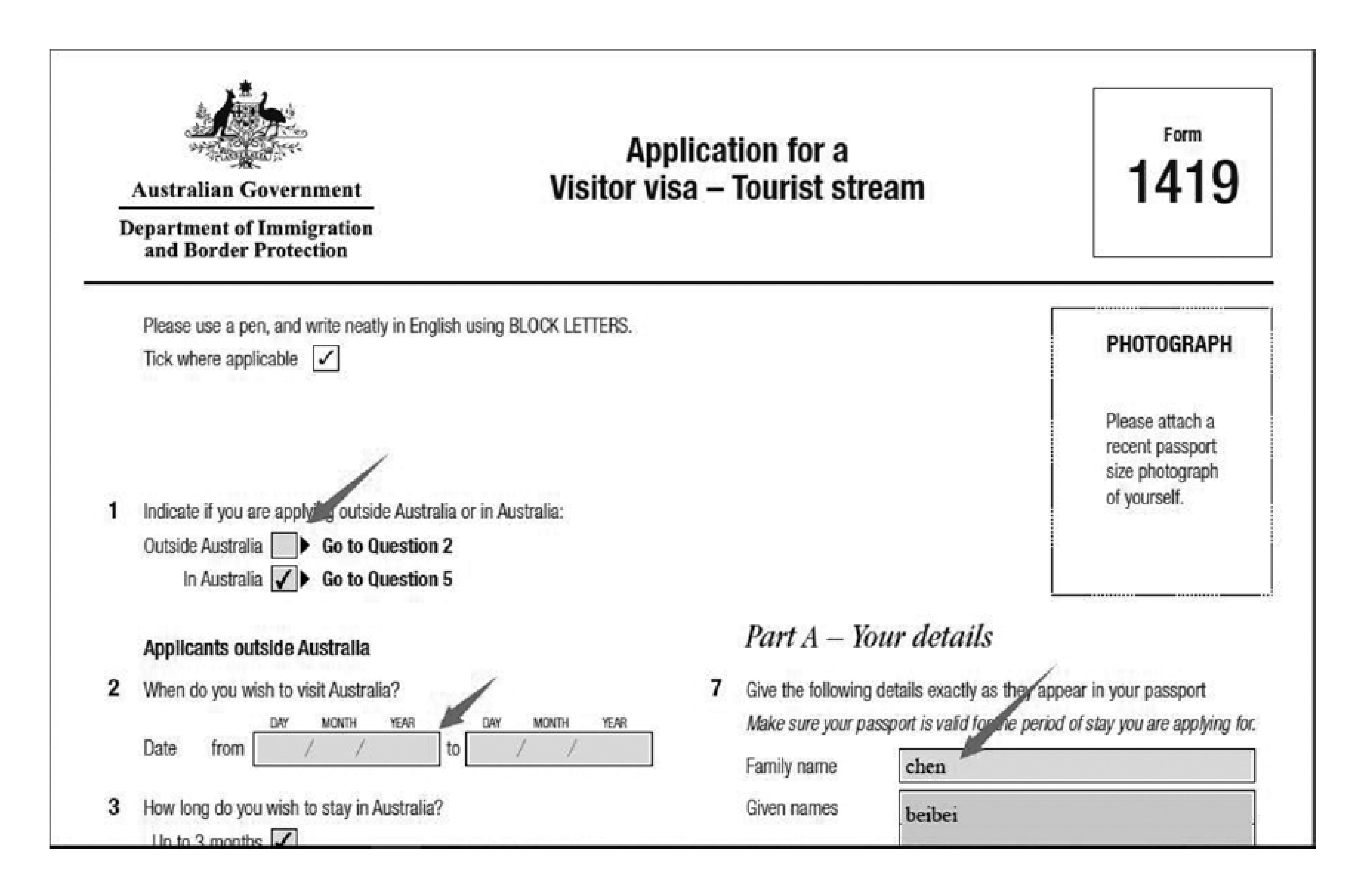
**公司薪酬满意度调查问卷

调查问卷说明：
本调查问卷共有48个问题，问题采用单项选择的方式，简明扼要并易于回答。
你可以匿名填写此份调查表。
本调查问卷的保密级为A级，任何信息都将严格受到保密，所以你可以放心做答。
当有超过50%的题目不做回答时，本问卷将做无效处理。
请你按实际情况作答，否则将影响调查结果。

你的姓名： 所在部门： 你的职位： 入职年限：
你的年龄： 性别： 学历程度：

题目	A	B	C	D	E	选择	如果选择D或E，请写明原因理由或感受
1. 你对自己努力付出与工资回报二者公平性的感受是	完全公平	基本公平	不确定	不公平	非常不公平		
2. 以自己的资历，你对自己的工资收入	非常满意	比较满意	不确定	不满意	非常不满意		
3. 领到工资时，你的感受是	比较愉快	比较开心	不确定	有些失落	心情很糟糕		
4. 你的努力工作在工资中有明显的回报吗？	一定有	可能有	不确定	没有	完全没有		
5. 和其他同职位的人相比，自己的工资	非常高	较高	不确定	较低	非常低		
6. 你觉得目前的工资就是你个人价值的体现吗？	肯定是	应该是	不确定	不是	绝对不是		
7. 你对目前公司薪酬制度科学性的评价是	非常科学	比较科学	不确定	不够科学	非常不科学		

图 2-8　不同版本问卷



Australian Government
Department of Immigration and Border Protection

Application for a Visitor visa – Tourist stream

Form 1419

Please use a pen, and write neatly in English using BLOCK LETTERS.
Tick where applicable ✓

PHOTOGRAPH
Please attach a recent passport size photograph of yourself.

1 Indicate if you are applying outside Australia or in Australia:
Outside Australia ☐ ▶ Go to Question 2
In Australia ☑ ▶ Go to Question 5

Applicants outside Australia

2 When do you wish to visit Australia?
Date from DAY / MONTH / YEAR to DAY / MONTH / YEAR

3 How long do you wish to stay in Australia?
Up to 3 months ☑

Part A – Your details

7 Give the following details exactly as they appear in your passport
Make sure your passport is valid for the period of stay you are applying for.
Family name: chen
Given names: beibei

图 2-9　可填写的 PDF 问卷

老梁：哦，想起来了，当时填的旅游申请表的确是一个 PDF 文件，里面有些内容的确可以输入文字或者进行选择，那时候还觉得挺神奇呢。不过，在 PDF 中设计这种可编辑的内容有什么意义呢？为什么不直接用

Word 或 Excel 来填写呢？如果用 PDF 文件来填写，提取数据的时候也会有大量的复制、粘贴操作啊。

Miss 陈：主要有三个原因。

（1）防止改变整个文档的设计和格式。PDF 文档整体不可编辑，文档的内容、排版、格式是固定的，这样能最大限度防止别人改动文档，保持文档的原始外貌，打印出来的文档排版和格式都是统一的。Word 和 Excel 格式的文档如果不做特别限定，整体都是可以编辑的，格式和排版都不受控制，往往导致回收的问卷样式各异。

（2）限制填写的内容和格式。PDF 文档中添加的可编辑元素，可以提供下拉菜单限制选择，可以限定文本框中只能填入数字或者日期、限定填写字数、限定单选或多选，通过这些方式可以最大限度地规范填报的信息，节省后期数据整理的时间。

（3）快速收集填写的数据。PDF 文档中填写的内容可以一次性批量导出到 Excel，并且能保证导出的格式规范、统一。用这种方式汇总数据准确、快速，比 Excel 或者 Word 的复制、粘贴操作强了很多，能大量节省数据整理的时间。

老梁：Word 确实不好控制版面和格式，每次发下去，等收回来的时候格式都被改得乱七八糟，填写的数据也很难汇总，复制、粘贴累死人。但是，PDF 文档中的这些可以编辑的区域是怎么做出来的，又如何快速地收集数据呢？

Miss 陈：其实做起来很简单，这需要用到 Acrobat 软件。你可能知道，PDF 是 Adobe 公司设计发明的一种跨平台的文件类型。查看 PDF 的软件是 Acrobat Reader，而编辑 PDF 的软件就是 Acrobat。它们都是 Adobe 公司的产品。

老梁：Adobe 公司我听说过，大名鼎鼎的 photoshop 就是 Adobe 公司的产品。Acrobat Reader 也在用，但是 Acrobat 就没用过。

Miss 陈：要编辑 PDF 就要用到 Acrobat。装好 Acrobat 之后，就可以制作可填写内容的 PDF 文件了。简单来说，只需要两步即可。

第一步：在 Word 中设计问卷，保存为 PDF 格式。在 Word 中设计问卷，并排好版。排版要尽量规范，需要别人填写的内容加一条下划线，需要单选的在选项前添加"○"，需要复选的在选项前添加"□"。然后将文件另存为 PDF 格式文档。Word 2007 及以上版本都可以将文档直接保存为 PDF 格式。

第二步：在 Acrobat 中打开 PDF 文档，使用表单工具创建表单。Acrobat 将自动识别文件中需要填写的内容，如各种下划线，○、□等符号，自动将它们转化为可以填写的表单域。Acrobat 中的表单域如图 2-10 所示。



图 2-10　Acrobat 中的表单域

经过以上两步，一个简单的可以填写的 PDF 文档就制作完成了。看看效果吧，如图 2-11 所示。

老梁：看上去很方便啊！用 Word 设计好问卷并另存为 PDF 格式，然后在 Acrobat 中利用表单工具就可以自动生成可填写的 PDF 文档。

Miss 陈：是的，在实际应用的时候，还可以添加一些小工具来规范输入的内容，这些小工具在 Acrobat 中叫作表单域。包括文本域、复选框、单选钮、列表框、下拉菜单、按钮、数字签名、条形码，等等。灵活运用这些表单域，就能制作满足我们需要的 PDF 文档了。

老梁：原来 PDF 还可以添加表单域啊。这些表单域看上去很像 html 网页中的控件，Word 的开发工具中也有类似的控件，倒不算陌生。

**公司薪酬满意度调查问卷

调查问卷说明：

- 本调查问卷共有 48 个问题，问题采用单项选择的方式，简明扼要并易于回答。
- 你可以匿名填写此份调查表。
- 本调查问卷的保密级为 A 级，任何信息都将严格受到保密，所以你可以放心做答。
- 当有超过 50%的题目不做回答时，本问卷将做无效处理。
- 请你按实际情况作答，否则将影响调查结果。

你的姓名：张三　　所在部门：财务部

你的职位：主管　　入职年限：3　　你的年龄：27

性　　别：男　　学历程度：本科

1. 你对自己努力付出与工资回报二者公平性的感受是

○ 完全公平

◉ 基本公平

○ 不确定

○ 不公平

○ 非常不公平

如果选择最后两项，请写明简要理由或感受：________。

图 2-11　可填写的薪酬满意度调查问卷

经理，您刚才说可以通过 PDF 快速收集填写的数据，这是真的吗？

Miss 陈：是真的。当我们将 PDF 问卷回收后，可将所有回收的 PDF 文件放到一个文件夹中，然后打开 Acrobat，用表单选项中的“合并数据文件到电子表格”选项，就可以一次性批量将数百乃至上千个 PDF 文档中的填写数据输出到一个 Excel 文件中。

在输出的 Excel 文件中，第一行是表单域的名称，从第二行开始就是每个文件填写的内容，数据排量相当规范，与数据库中的数据表类似。

老梁：太方便了，不得不说，用 PDF 制作问卷来收集数据实在是一

个非常方便快捷的方式啊，以前怎么没发现这个工具呢？

2.3.2 利用互联网、手机微信进行问卷调查

Miss 陈：如果问卷内容的涉密程度不是很高，还可以利用互联网进行问卷调查。

老梁：您是说我们建立一个服务器，然后把问卷发布到网上吗？

Miss 陈：不是，这样成本会很高。网上已经有不少在线问卷发布网站，提供了平台，我们只需要把问卷导入这类平台，就可以发布问卷了。这类平台通常收费较低，有些甚至是免费的。

但是由于问卷内容和填报的数据都放到了网上，信息泄露的概率升高，所以通常仅适用于涉密程度不是很高的问卷调查。

老梁：还有免费的啊？会不会功能上有限制呢？比如有些题型不能添加，或者数据达到一定程度就要收费，等等。

Miss 陈：这类网站形形色色，你可以搜索一下，逐个试试看，看看各有什么特点。这里推荐使用问卷网（www. wenjuan. com），理由如下。

（1）基本功能完全免费，不限制问卷数量、题型、答题人数，无广告。

（2）统计分析、报表功能完善，原始数据可以随时下载。

（3）问卷类型多样，排版简洁、干净。

（4）与手机微信衔接很好，可以通过微信发布、填写问卷，问卷提交后可以给微信发送实时提醒，如图 2-12 所示。

老梁：还能和微信联系起来，太好了，这样员工不用打开电脑，直接在手机上就可以填写问卷了，这种形式符合现在移动互联网的发展形势，很有创意啊。我得赶快去试试。

●○○○○ 中国电信　23:22

返回　**公司薪酬满意度调查问卷　•••

**公司薪酬满意度调查问卷

调查问卷说明：

本调查问卷共有48个问题，问题采用单项选择的方式，简明扼要并易于回答。

你可以匿名填写此份调查表。

本调查问卷的保密级为A级，任何信息都将严格受到保密，所以你可以放心做答。

当有超过50%的题目不做回答时，本问卷将做无效处理。

请你按实际情况作答，否则将影响调查结果。

矩阵填空题

你的姓名

请填写

所在部门

请填写

你的职位

图 2-12　手机微信填写薪酬满意度调查问卷

2.4 整理数据

2.4.1 关于一维表

老梁：经理，收集了数据之后，是不是就可以进行数据分析了呢？

Miss 陈：这需要看数据的质量。如果数据质量高的话，就可以进行分析；如果数据质量不高的话，就需要进行数据整理，也叫作数据清洗。

就好比我们做菜，蔬菜可能含有农药要用水漂洗多次，肉需要切成肉丝或肉丁，鱼要去鳞清洗肚腹等，这是将原材料加工成可以烹饪的形态，然后才进行烹饪，而不是直接就开始烹饪，否则做出来的菜谁也不敢吃。数据整理就是对数据进行清洗的过程，清洗后才能进行数据分析。

遗憾的是，多数时候收集到的数据质量都不太高，或多或少有些问题，所以需要进行数据整理，而且这个过程是数据分析中耗时最长的。

老梁：那么要怎样进行数据整理呢？

Miss 陈：首先我们讲一个数据整理中比较常见的问题。你看下面的两张表，表中数据是员工绩效考核成绩，表 2-1 和表 2-2 的数据有什么不同？

表 2-1　数据整理样表一

部门	性别	绩效总分	情绪总分	适应总分
⊟1	1	11.45	12.05	10.00
	2	10.91	10.82	7.91
⊟2	1	10.29	10.29	9.00
	2	10.29	10.06	10.00
⊟3	1	15.25	12.90	12.00
	2	12.06	12.31	9.38
总计		11.89	11.54	9.93

表 2-2　数据整理样表二

ID	员工编号	性别	部门	绩效总分	适应总分	情绪总分
1	1	1	1	12.00	11.00	12.00
2	2	1	3	13.00	10.00	12.00
3	3	1	1	20.00	10.00	14.00
4	4	2	2	8.00	12.00	8.00
5	5	2	3	11.00	12.00	12.00
6	6	2	1	11.00	11.00	10.00
7	7	2	3	14.00	8.00	11.00
8	8	2	1	11.00	10.00	13.00
9	9	2	3	6.00	9.00	10.00
10	10	2	2	6.00	6.00	9.00
11	11	1	1	7.00	10.00	7.00

老梁：我看看，表 2-1 好像是汇总表，是我们常用来做报表的样式；表 2-2 则是一排排数据，像是数据库中的数据表。

Miss 陈：是的。平常我们用表 2-1 的表格较多，这种表格称为二维表。就是横排和纵列分别代表不同的维度，甚至多个维度。比如表 2-1 中，横排是考核分数的类型，纵列是部门和性别。绝大多数情况下，这种二维表不适合做数据分析。

老梁：您的意思是像表 2-2 样式的表格才适合进行数据分析吗？

Miss 陈：是的，像表 2-2 这种样式的数据我们称之为一维表，第一行代表了分析的维度，从第二行开始就是一条条数据。在不同的知识领域，一维表的叫法不同，比如在数据库中，一维表叫作数据表，第一行是字段，第二行开始叫作记录；在数据分析中，第一行叫作变量，从第二行开始叫作观测值。

老梁：原来如此，我以前接触过数据库，对这种形式的数据还是有些了解的，但不知道数据分析需要这种形式。

Miss 陈：数据分析领域绝大多数的算法，都是基于一维表进行的，所以如果我们要进行数据分析，但手头上只有二维表，就必须将二维表转换为一维表，才能进行数据分析。

老梁：这个很重要，我们很多报表的数据都是二维表，照这样看都不能进行数据分析，要转换成一维表才行。但是怎样将二维表转换成一维表呢？如果将一个个数据拆开再组合起来，要花不少时间呢，还容易出错。

Miss 陈：这种二维表转一维表的数据转换可以用 Excel 来做，整个操作方便又快捷。以表 2-3 的数据为例。

表 2-3 是性别和三个评价维度组成的二维表，其中绩效总分、情绪总分、适应总分都是对员工绩效的测评结果，可以合并为一个变量，但这里分成了三个变量。下面我们看看如何用 Excel 来将这种二维表转换为一维表。

表 2-3 待转换的二维表

性别	绩效总分	情绪总分	适应总分
男	11.45	12.05	10.00
女	10.91	10.82	7.91
男	10.29	10.29	9.00
女	10.29	10.06	10.00
男	15.25	12.90	12.00
女	12.06	12.31	9.38

（1）打开 Excel，先按“Alt＋D”键，然后再按“P”键，在弹出的对话框中打开“数据透视表和数据透视图向导”对话框，选中“多重合并计算数据区域”选项，如图 2-13 所示。



图 2-13 数据透视表和数据透视图向导

（2）选中“创建单页字段”选项，单击“下一页”按钮，如图 2-14 所示。

（3）在“选择区域”中选中待转换的二维表，单击“添加”按钮，如图 2-15 所示。

（4）单击“新建工作表”按钮，再单击“完成”按钮，此时会生成一个数

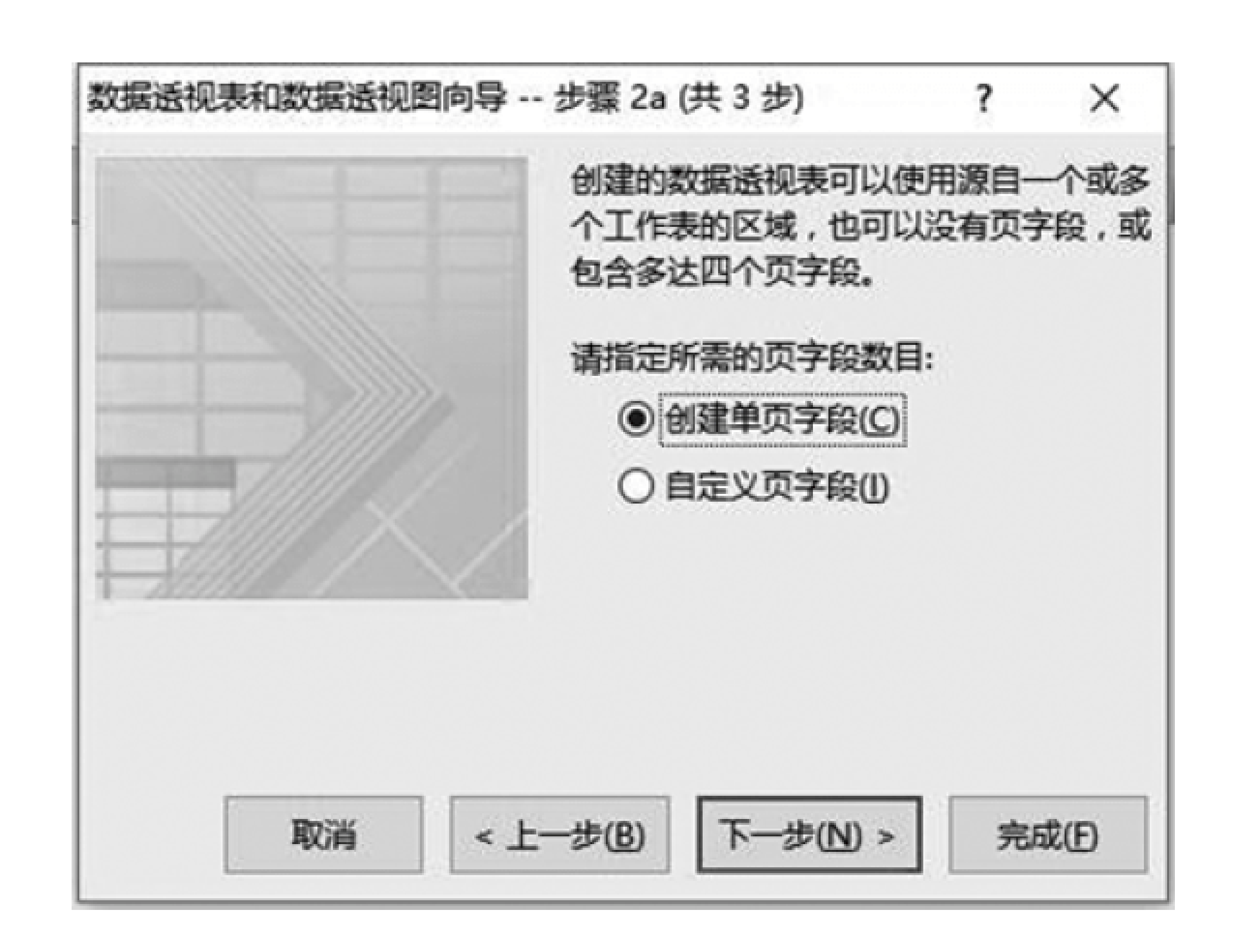


图 2-14 创建单页字段



图 2-15 添加待转换的数据区域

据透视表，如图 2-16 所示。

（5）双击数据透视表的最后一个单元格，Excel 会自动创建一个新的工作表，新的工作表就是转换后的一维表，具体见表 2-4。



图 2-16 生成数据透视表

表 2-4 转换后的一维表

行	列	值	页1
男	绩效总分	11.454 5	项1
男	绩效总分	10.285 7	项1
男	绩效总分	15.25	项1
男	情绪总分	12.045 5	项1
男	情绪总分	10.285 7	项1
男	情绪总分	12.9	项1
男	适应总分	10	项1
男	适应总分	9	项1
男	适应总分	12	项1
女	绩效总分	10.909 1	项1
女	绩效总分	10.294 1	项1
女	绩效总分	12.062 5	项1
女	情绪总分	10.818 2	项1
女	情绪总分	10.058 8	项1
女	情绪总分	12.312 5	项1
女	适应总分	7.909 09	项1
女	适应总分	10	项1
女	适应总分	9.375	项1

老梁：原来 Excel 还有这种隐藏的功能啊！

Miss 陈：Excel 把二维表转换为一维表时利用了数据透视表的功能，转换过程显得简单、直观、快捷，所以遇到这种情况用 Excel 最方便。

当把数据转换为一维表后，就可以导入 R 语言进行数据的整理和分析了。

2.4.2　处理缺失值

Miss 陈：进行数据整理时最常碰到的，最令人头疼的事情是出现数据缺失，就是数据不完整，出现了空值。这种缺失会影响数据分析的效果，导致分析结论出现错误。

老梁：是啊，辛辛苦苦收集来的数据，发现这里缺少数据，那里缺少数据，很头疼。这种缺失情况是怎么造成的呢？

Miss 陈：造成数据缺失的原因是多种多样的，总体来说分为机械原因和人为原因。机械原因是指由于硬件原因导致数据收集或保存失败而造成的数据缺失。比如数据存储失败，存储器损坏，机械故障导致某段时间的数据未能收集。而人为原因是人的主观失误、历史局限或有意隐瞒造成的数据缺失，比如在问卷调研中填报人拒绝透露相关问题的答案，或者回答的问题是无效的、是谎言，再比如数据录入人员在录入数据时失误，漏录了数据，等等。这些原因都会造成数据缺失。

老梁：那怎么检查数据是否有缺失呢？

Miss 陈：最简单的办法是打开数据看看。

老梁：哎呀，糊涂了，直接打开看不就知道了吗？哈哈。

Miss 陈：如果数据量大的的话，直接看就比较花时间了。在数据量大的情况下，要直观了解数据缺失情况，可以用 R 语言中的 VIM 包函数 aggr 来查看。以我们公司在应届大学毕业生招聘时的测评数据为例，数据见表 2-5（数据较多，只显示了其中一部分）。

其缺失值情况经分析绘制成图形，如图 2-17 所示。

表 2-5　应届大学毕业生招聘时的测评数据

序号	姓名	言语理解	数学	逻辑	常识	成就导向	抗压能力	灵活性	影响性	支配性	外向性
1	梁**	8.00	15.50	22.60	3.50	7.96	4.46	5.18	6.87	7.84	6.28
2	李**	9.00	11.50	17.80	4.20	6.01	4.46	3.54	4.21	5.63	5.14
3	傅**	10.00	9.50	12.00	4.90	5.62	6.86	5.72	6.87	6.37	6.66
4	叶**	9.00	12.50	15.90	4.90		4.46	6.27	7.53	8.57	6.28
5	韩**	15.00	24.80	25.60	10.80	4.84	5.06	2.45	4.21	4.90	5.52
6	骆**	15.00	20.50	21.90	9.60	5.23	6.26	5.18	4.87	6.73	5.52
7	姚**	10.50	22.50	24.50	10.80	5.38	7.11	4.63	4.87	4.90	4.76
8	余**	13.50	22.50	17.90	10.80	7.57	5.96	4.09	5.54	7.84	5.14
9	姚**	15.00	22.50	16.50	9.60	5.23	5.36	7.36	6.20	5.26	5.90
10	蔡**	19.50	28.50	29.00	12.00	7.18	5.36	7.36	8.20	7.47	6.28
11	苏**	18.00	24.50	37.00	12.00	3.62	3.56	8.45	2.88	2.69	3.61
12	林**	13.50	22.80	24.20	9.60	3.67	5.81	2.45	4.87	6.37	3.99
13	黄**	10.50	20.80	17.10	8.40	2.89	5.36	3.00	5.54	5.26	7.42
14	王**	8.65	11.82	18.39	5.43	4.47	4.47	4.47	4.47	4.47	4.47
15	卢**	9.00	10.50	24.80	7.00	4.45	2.96	6.27	1.55	5.63	2.09
16	高**	13.50	10.80	24.80	7.20	4.84	6.46	4.09	2.21	3.06	7.04

图 2-17 的左边是缺失值在各个变量中的占比情况，右边是各个变量中的缺失值分布情况，黑色代表有缺失值。通过图形就能比较直观地看到缺失值的情况了。

处理缺失值的 R 语句如下：

```
library(VIM)
#读取数据
d<-read.csv("第二章/毕业生数据.csv")
# 检查缺失值情况
aggr(d)
```

老梁：有这么多缺失值啊，该怎么办呢?

Miss 陈：别担心，缺失值是比较常见的，只要比例不超过总数据量的 10%，影响也不见得很大。不过咱们还是要对这些缺失值进行处理。

对于缺失值的处理，常用的方法有：直接删除法、均值插补法、同类值插补法、极大似然估计法、多重插补法。

比较简单的是直接删除法，就是直接将有缺失值的那行数据删除，但是会带来数据信息的流失。如果数据量很大，并且缺失值不多的时候，用直接删除法不失为一种简单、经济、快速的方法。

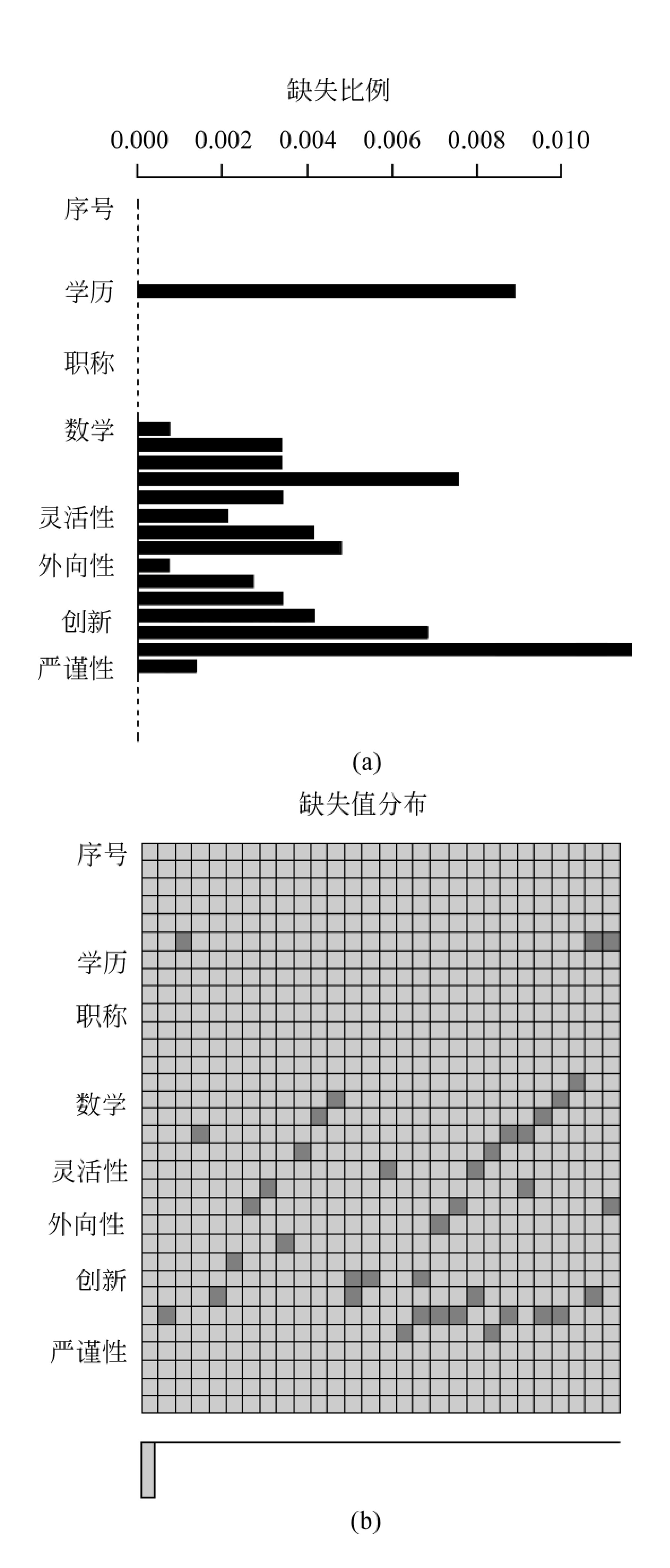


图 2-17　应届大学毕业生招聘时测评数据的缺失情况

比较常用的方法是多重插补法。多重插补的思想源于贝叶斯估计，该理论认为待插补的值是随机的，它的值来自已观测到的值。多重插补通过变量间关系来预测缺失数据，利用蒙特卡罗方法生成多个完整数据集，再对这些数据集分别进行分析，最后对这些分析结果进行汇总处理。在 R 语言中是使用 mice 包中的 mice 函数。

采用多层插补法的 R 语句如下：

```
library(mice)
# 读取数据
d<- read.csv("第二章/毕业生数据.csv")
# 采用多重插补法填补缺失数据
d1<- mice(d)
```

上述语句中，已经通过 mice 函数将毕业生测评数据的缺失值进行了多重插补运算，默认生成 5 组插补值，然后存储到变量 d1 中，之后的各种分析用 d1 来进行即可。

老梁：多重插补法虽然比较复杂，但计算过程交给函数去处理，倒也省事啊。

2.4.3 处理重复数据

老梁：经理，我经常在工作中遇到重复数据，这种情况是否也要处理呢？比如前两天我们收集各个分公司培训主管的个人数据，也不知道上报的人员是怎么搞的，报上来的数据有些是重复的，拿到这种数据要怎么处理呢？具体数据见表 2-6。

Miss 陈：重复数据也是常见的错误数据类型，需要进行清洗，删除多余的重复数据。在 R 语言中可以用 unique 函数进行去除重复数据的操作。清洗后的结果见表 2-7。

表 2-6　各分公司培训主管基本信息数据(清洗前)

姓　名	性别	岗　位	层级	职业资格	职　称
李平川	女	高级人力资源主管	七岗	人力资源管理师	
周雅松	女	高级人力资源业务员	十一岗	助理人力资源管理师	
吴雷	男	人力资源主办	十岗	助理人力资源管理师	助理经济师
韩磊	女	培训主管	十岗		
欧阳志远	女	高级人力资源主办	九岗	高级人力资源管理师	助理工程师
李力持	男	职能部门室副经理	八岗	高级人力资源管理师	工程师
周晓薇	女	高级主管	七岗	人力资源管理师	高级经济师
陈宇宙	女	高级人力主管	七岗	人力资源管理师、内部培训师	高级经济师
罗敏	女	人力资源主办	十岗		
张元	男	高级人力资源主办	九岗	物业管理师	助理经济师
荣波	男	高级人力资源主办	九岗		
王一帅	男	培训招聘主管	八岗		
王南溪	女	总工室技术管理室经理	七岗		工程师
张元	男	高级人力资源主办	九岗	物业管理师	助理经济师
杨一	女	高级人力资源管理员	十一岗		助理工程师
李敏	女	高级人力资源业务员	十一岗	高级人力资源管理师	
杨单博	女	综合室副经理	八岗	高级人力资源管理师	高级工程师
杨一	女	高级人力资源管理员	十一岗		助理工程师
范丁	女	培训主管	八岗	计算机中级	
孟津	男	培训主管	九岗	概预算员	工程师
郑波	女	辅助办事员	十五岗	助理人力资源管理师	
朱进权	女	高级人力资源业务员	十岗	助理人力资源管理师	助理工程师
李锐	女	高级人力资源主办	九岗	高级人力资源管理师	经济师
何磊	女	人力资源高级主管	七岗	高级人力资源管理师	经济师
白钢	女	高级人力资源主办	九岗	高级人力资源管理师、劳动关系协调师	助理工程师
夏琳香	女	办事员	十一岗		

续表

姓　名	性别	岗　位	层级	职业资格	职　称
何磊	女	人力资源高级主管	七岗	高级人力资源管理师	经济师
李锐	女	高级人力资源主办	九岗	高级人力资源管理师	经济师
钟慧	女	一级综合业务员	十一岗	教师资格证书	幼儿园一级教师
陈侠	女	高级人力资源业务员	十一岗		
李梅	女	高级人力资源主管	七岗	高级人力资源管理师	高级经济师
孟丽	女	高级人力资源主管	七岗	高级人力资源管理师	经济师
余水	男	人力资源部培训室经理	七岗	监理工程师、信息系统监理师、评标专家、概预算、安全工程师、PMP	高级工程师
吴春荣	女	高级人力主办	九岗	高级人力资源管理师	经济师
雷雨薇	男	高级设计师	九岗		
郑启荣	女	人力资源主办	十岗	人力资源管理师	
余水	男	人力资源部培训室经理	七岗	监理工程师、信息系统监理师、评标专家、概预算、安全工程师、PMP	高级工程师

表 2-7　各分公司培训主管基本信息数据(清洗后)

姓　名	性别	岗　位	层级	职业资格	职　称
李平川	女	高级人力资源主管	七岗	人力资源管理师	
周雅松	女	高级人力资源业务员	十一岗	助理人力资源管理师	
吴雷	男	人力资源主办	十岗	助理人力资源管理师	助理经济师
韩磊	女	培训主管	十岗		
欧阳志远	女	高级人力资源主办	九岗	高级人力资源管理师	助理工程师
李力持	男	职能部门室副经理	八岗	高级人力资源管理师	工程师
周晓薇	女	高级主管	七岗	人力资源管理师	高级经济师

续表

姓　名	性别	岗　位	层级	职 业 资 格	职　称
陈宇宙	女	高级人力主管	七岗	人力资源管理师、内部培训师	高级经济师
罗敏	女	人力资源主办	十岗		
张元	男	高级人力资源主办	九岗	物业管理师	助理经济师
荣波	男	高级人力资源主办	九岗		
王一帅	男	培训招聘主管	八岗		
王南溪	女	总工室技术管理室室经理	七岗		工程师
杨一	女	高级人力资源管理员	十一岗		助理工程师
李敏	女	高级人力资源业务员	十一岗	高级人力资源管理师	
杨单博	女	综合室副经理	八岗	高级人力资源管理师	高级工程师
范丁	女	培训主管	八岗	计算机中级	
孟津	男	培训主管	九岗	概预算员	工程师
郑波	女	辅助办事员	十五岗	助理人力资源管理师	
朱进权	女	高级人力资源业务员	十岗	助理人力资源管理师	助理工程师
李锐	女	高级人力资源主办	九岗	高级人力资源管理师	经济师
何磊	女	人力资源高级主管	七岗	高级人力资源管理师	经济师
白钢	女	高级人力资源主办	九岗	高级人力资源管理师、劳动关系协调师	助理工程师
夏琳香	女	办事员	十一岗		
钟慧	女	一级综合业务员	十一岗	教师资格证书	幼儿园一级教师
陈侠	女	高级人力资源业务员	十一岗		
李梅	女	高级人力资源主管	七岗	高级人力资源管理师	高级经济师
孟丽	女	高级人力资源主管	七岗	高级人力资源管理师	经济师

续表

姓　名	性别	岗　位	层级	职业资格	职　称
余水	男	人力资源部培训室经理	七岗	监理工程师、信息系统监理师、评标专家、概预算、安全工程师、PMP	高级工程师
吴春荣	女	高级人力主办	九岗	高级人力资源管理师	经济师
雷雨薇	男	高级设计师	九岗		
郑启荣	女	人力资源主办	十岗	人力资源管理师	

处理重复数据的 R 语句如下：

```
d<-read.csv("第二章/培训人员信息.csv")
# 剔除重复数据
d<-unique(d)
# 保存剔除重复值后的数据
write.csv(d,"第二章/培训人员信息(去重复).csv")
```

老梁：真方便，原来 R 语言中剔除重复数据用一个函数就完成了啊。

2.4.4 数据分组

Miss 陈：老梁，再问你一个问题，你遇到过需要对数据进行分组的情况吗？

老梁：经常遇到，比如按年龄段分组，按单位分组，按薪酬等级分组等，分组之后就可以按组别进行统计。

Miss 陈：嗯，分组也是进行数据整理的一个重要环节。在数据分析领域，这种分组叫作分类，是将连续数据转换为类别(因子)数据的过程。举个例子，就用刚才各分公司培训人员数据。我们知道每个人都有年龄，现在我们拟对年龄进行分组。根据观察，最小的 20 岁，最大的 45 岁，可以分为三个组别，分别是：①30 岁及以下；②大于 30 岁小于等于 40 岁；

③大于 40 岁小于等于 50 岁。我们就按照这个规则对数据进行分组吧，分组后的数据见表 2-8，分组结果见“年龄组”字段。

表 2-8　各分公司培训主管基本信息数据（分组数据）

姓　名	性别	岗　位	层　级	职业资格	职　称	年龄（岁）	年龄组
李平川	女	高级人力资源主管	七岗	人力资源管理师		39	30～40 岁
周雅松	女	高级人力资源业务员	十一岗	助理人力资源管理师		30	30 岁及以下
吴雷	男	人力资源主办	十岗	助理人力资源管理师	助理经济师	39	30～40 岁
韩磊	女	培训主管	十岗			40	30～40 岁
欧阳志远	女	高级人力资源主办	九岗	高级人力资源管理师	助理工程师	45	40～50 岁
李力持	男	职能部门室副经理	八岗	高级人力资源管理师	工程师	22	30 岁及以下
周晓薇	女	高级主管	七岗	人力资源管理师	高级经济师	35	30～40 岁
陈宇宙	女	高级人力主管	七岗	人力资源管理师、内部培训师	高级经济师	33	30～40 岁
罗敏	女	人力资源主办	十岗			40	30～40 岁
张元	男	高级人力资源主办	九岗	物业管理师	助理经济师	26	30 岁及以下
荣波	男	高级人力资源主办	九岗			39	30～40 岁
王一帅	男	培训招聘主管	八岗			33	30～40 岁
王南溪	女	总工室技术管理室经理	七岗		工程师	36	30～40 岁

续表

姓　名	性别	岗　位	层　级	职业资格	职　称	年龄（岁）	年龄组
张元	男	高级人力资源主办	九岗	物业管理师	助理经济师	22	30 岁及以下
杨一	女	高级人力资源管理员	十一岗		助理工程师	21	30 岁及以下
李敏	女	高级人力资源业务员	十一岗	高级人力资源管理师		37	30～40 岁
杨单博	女	综合室副经理	八岗	高级人力资源管理师	高级工程师	42	40～50 岁
杨一	女	高级人力资源管理员	十一岗		助理工程师	33	30～40 岁
范丁	女	培训主管	八岗	计算机中级		40	30～40 岁
孟津	男	培训主管	九岗	概预算员	工程师	43	40～50 岁
郑波	女	辅助办事员	十五岗	助理人力资源管理师		43	40～50 岁
朱进权	女	高级人力资源业务员	十岗	助理人力资源管理师	助理工程师	30	30 岁及以下
李锐	女	高级人力资源主办	九岗	高级人力资源管理师	经济师	43	40～50 岁
何磊	女	人力资源高级主管	七岗	高级人力资源管理师	经济师	28	30 岁及以下
白钢	女	高级人力资源主办	九岗	高级人力资源管理师、劳动关系协调师	助理工程师	40	30～40 岁
夏琳香	女	办事员	十一岗			23	30 岁及以下
何磊	女	人力资源高级主管	七岗	高级人力资源管理师	经济师	25	30 岁及以下
李锐	女	高级人力资源主办	九岗	高级人力资源管理师	经济师	32	30～40 岁

续表

姓　名	性别	岗　位	层　级	职业资格	职　称	年龄(岁)	年龄组
钟慧	女	一级综合业务员	十一岗	教师资格证书	幼儿园一级教师	27	30 岁及以下
陈侠	女	高级人力资源业务员	十一岗			29	30 岁及以下
李梅	女	高级人力资源主管	七岗	高级人力资源管理师	高级经济师	35	30～40 岁
孟丽	女	高级人力资源主管	七岗	高级人力资源管理师	经济师	20	30 岁及以下
余水	男	人力资源部培训室经理	七岗	监理工程师、信息系统监理师、评标专家、概预算、安全工程师、PMP	高级工程师	33	30～40 岁
吴春荣	女	高级人力主办	九岗	高级人力资源管理师	经济师	40	30～40 岁
雷雨薇	男	高级设计师	九岗			31	30～40 岁
郑启荣	女	人力资源主办	十岗	人力资源管理师		24	30 岁及以下
余水	男	人力资源部培训室经理	七岗	监理工程师、信息系统监理师、评标专家、概预算、安全工程师、PMP	高级工程师	31	30～40 岁

进行数据分组的 R 语句如下：

```
d<-read.csv("第二章/培训人员信息.csv")
#数据分组(按年龄分组)
d$年龄组<-cut(d$年龄,breaks=c(0,30,40,50),labels=c("30 岁及以
```

```
下","30～40岁","40～50岁"))
#保存分组后的数据
write.csv(d,"第二章/培训人员信息(分组).csv")
```

老梁：明白了。不过在数据分析中将原始的连续数据转换为分组数据有什么特别用意吗？

Miss陈：是的，分组后有许多好处呢，最直接的好处就是方便进行分组统计，可以计算每组数据的均值、方差等；还可以进行组间对比分析，研究不同组之间的差异情况，比如可以分析不同年龄组人员之间的绩效是否有差异，哪个年龄组绩效最高，哪个最低；更重要的是分组数据还可以通过一些算法进行预测，如通过判别分析、机器学习等算法可以建立分析模型，用模型对未知情况进行预测。你还记得前段时间网上流行的传照片测年龄的游戏吗？

老梁：记得记得，就是把照片传到一个网站上，该网站就能显示照片中每个人的年龄。我们都试了一下，还挺准呢。现在有些手机在照相时也可以显示年龄，挺神奇。这是怎么做到的呢？难道和数据分组有关系吗？

Miss陈：其实是运用了分类算法，年龄就是分组，只不过这个分组比较细，按1岁来分组。在收集了大量的人脸信息数据和年龄数据后，就可以通过统计分析软件，用分类算法建立统计模型，模型建立之后就可以根据照片中的人脸信息计算年龄了。

老梁：原来如此。

2.4.5 生成新数据

Miss陈：还有些时候，我们想根据一列数据生成另一列数据，用新生成的数据来进行分析，这时候就需要通过计算产生新变量。

老梁：嗯，常碰到这种情况呢。我上个月进行员工薪酬分析的时候，

想要分析员工薪酬和市场薪酬之间的差距，就通过计算直接得出员工薪酬和市场薪酬之间的差距，然后再进行分析。

Miss 陈：对于通过计算生成新变量的情况，只要我们确定了计算规则，剩下的就好办了。还是拿刚才培训人员的数据来举例吧，如我们公司员工的平均年龄是 29 岁，现在想知道每个人的年龄和平均年龄之间的差距是多少。这时，我们需要生成一个新的变量，用来保存年龄差距，可以将这个变量命名为“与平均年龄之差”。这个新变量的计算规则比较简单，是员工年龄与平均年龄之差，通过计算，结果见表 2-9。

表 2-9　各分公司培训主管基本信息数据（生成新数据）

姓　名	性别	岗　位	层　级	职业资格	职　称	年龄（岁）	与平均年龄之差
李平川	女	高级人力资源主管	七岗	人力资源管理师		39	10
周雅松	女	高级人力资源业务员	十一岗	助理人力资源管理师		30	1
吴雷	男	人力资源主办	十岗	助理人力资源管理师	助理经济师	39	10
韩磊	女	培训主管	十岗			40	11
欧阳志远	女	高级人力资源主办	九岗	高级人力资源管理师	助理工程师	45	16
李力持	男	职能部门室副经理	八岗	高级人力资源管理师	工程师	22	−7
周晓薇	女	高级主管	七岗	人力资源管理师	高级经济师	35	6
陈宇宙	女	高级人力主管	七岗	人力资源管理师、内部培训师	高级经济师	33	4
罗敏	女	人力资源主办	十岗			40	11
张元	男	高级人力资源主办	九岗	物业管理师	助理经济师	26	−3

续表

姓 名	性别	岗 位	层 级	职业资格	职 称	年龄（岁）	与平均年龄之差
荣波	男	高级人力资源主办	九岗			39	10
王一帅	男	培训招聘主管	八岗			33	4
王南溪	女	总工室技术管理室经理	七岗		工程师	36	7
张元	男	高级人力资源主办	九岗	物业管理师	助理经济师	22	−7
杨一	女	高级人力资源管理员	十一岗		助理工程师	21	−8
李敏	女	高级人力资源业务员	十一岗	高级人力资源管理师		37	8
杨单博	女	综合室副经理	八岗	高级人力资源管理师	高级工程师	42	13
杨一	女	高级人力资源管理员	十一岗		助理工程师	33	4
范丁	女	培训主管	八岗	计算机中级		40	11
孟津	男	培训主管	九岗	概预算员	工程师	43	14
郑波	女	辅助办事员	十五岗	助理人力资源管理师		43	14
朱进权	女	高级人力资源业务员	十岗	助理人力资源管理师	助理工程师	30	1
李锐	女	高级人力资源主办	九岗	高级人力资源管理师	经济师	43	14
何磊	女	人力资源高级主管	七岗	高级人力资源管理师	经济师	28	−1
白钢	女	高级人力资源主办	九岗	高级人力资源管理师、劳动关系协调师	助理工程师	40	11

续表

姓　名	性别	岗　位	层　级	职业资格	职　称	年龄（岁）	与平均年龄之差
夏琳香	女	办事员	十一岗			23	−6
何磊	女	人力资源高级主管	七岗	高级人力资源管理师	经济师	25	−4
李锐	女	高级人力资源主办	九岗	高级人力资源管理师	经济师	32	3
钟慧	女	一级综合业务员	十一岗	教师资格证书	幼儿园一级教师	27	−2
陈侠	女	高级人力资源业务员	十一岗			29	0
李梅	女	高级人力资源主管	七岗	高级人力资源管理师	高级经济师	35	6
孟丽	女	高级人力资源主管	七岗	高级人力资源管理师	经济师	20	−9
余水	男	人力资源部培训室经理	七岗	监理工程师、信息系统监理师、评标专家、概预算、安全工程师、PMP	高级工程师	33	4
吴春荣	女	高级人力主办	九岗	高级人力资源管理师	经济师	40	11
雷雨薇	男	高级设计师	九岗			31	2
郑启荣	女	人力资源主办	十岗	人力资源管理师		24	−5
余水	男	人力资源部培训室经理	七岗	监理工程师、信息系统监理师、评标专家、概预算、安全工程师、PMP	高级工程师	31	2

生成新数据的 R 语句如下：

```
d<-read.csv("第二章/培训人员信息.csv")
#生成新变量
d$与平均年龄之差<-d$年龄- 29
#保存计算后的数据
write.csv(d,"第二章/培训人员信息(计算新变量).csv")
```

老梁：原来是这样生成新变量的，看上去只要知道了新变量的计算规则，就很容易操作了。

Miss 陈：是的。关于数据整理咱们就谈这么多。你明白了吧，数据整理涉及很多方面的内容。

老梁：是啊，没想到整理数据这么麻烦。

Miss 陈：其实数据整理还不止这些内容。

老梁：经理，还有什么技术，您给点提示吧。如果您没有时间，我们可以自己去找资料学习啊。

Miss 陈：在数据整理方面，还有一些技术我们没讲到，包括以下方面。

（1）数据抽样：如果数据量很大，导致数据分析速度很慢，可以考虑通过数据抽样的方法，抽取一部分数据作为样本，来代表总体进行分析。

（2）噪声处理：就是异常值的处理。有时候数据包含一些极端值、异常值，这些数据的存在会较大地影响数据分析、建模、预测的结果，可以通过噪声处理的技术剔除这些数据。

（3）数据集成：如果有多个相关联的数据表，比如有员工培训、薪酬、绩效考核三张表，都与员工相关，那么可以通过数据集成的技术将这三张表合并成一张表，进行分析，这有点儿像数据库中的数据表联结。

（4）数据标准化：将不同量纲的数据转换为量纲一致的数据，以避免因为量纲不同带来的分析误差。一般是将不同量纲的数据进行标准化或者归一化，转换为标准分或者[－1,1]之间的数值。

老梁：难怪数据整理要花很多时间，原来有这么多内容要处理，我得赶快去网上查找资料学习一下，免得连数据都整理不好，就更别提进行后续的分析了。还有，今后要求各个分公司报数据时一定要准确，一旦发现不准确的就退回重报。咱们得把数据整理的工作分摊给大家，尽量保证快速、高效地收集高质量的数据，减少数据整理的时间。

Miss 陈：老梁你果然是位经验丰富的人力资源管理人员啊，知道从管理方式入手改进数据质量。

第 3 章

员工年度需求预测

导语：传统的员工年度需求预测多采用自下而上的方法，由下属单位上报需求汇总而成；或者用经验法进行预测。这些方法存在预测精度不高，误差较大等问题。本章介绍模型法，即在收集与用工需求相关历史数据的基础上，通过建立回归模型，比较准确地预测公司下一年度的员工需求。

3.1 需求描述

某天，小肖来到 Miss 陈办公室汇报工作，提到公司明年的员工招聘计划。

小肖：经理，近期要制订明年的员工招聘计划了，但是我对明年需要招聘的新员工人数没有把握，您给指导指导吧。

Miss 陈：那你说说公司往年是怎样确定员工需求人数的。

小肖：以前我们用的方法比较简单，采取上报制度，通过层层上报，让下属单位上报需求人数，然后我们汇总需求，以此作为公司下一年的员工需求人数基数，制订招聘计划。至于下属单位是如何确定需求人数的，我们没有干预。

Miss 陈：你是不是觉得这种方式有问题呢？

小肖：是的，最明显的问题是需求人数不准确。最近几年下属单位上报的需求人数往往不够准确，甚至会出现比较大的偏差，给我们的招聘工作带来了困扰。

比如，各单位上报应届毕业生需求，通常在 9 月上报，实际招聘要持续到下一年四五月才结束，时间跨度比较大，在这个过程中各单位的毕业生需求会发生变化，但这种变化很晚才能反馈到我们这里。等到我们已经跟应届毕业生签订了三方协议后，用人单位突然告诉我们不需要招聘了，因为没有用工需求了。您想想这时候多郁闷啊，人都招聘好了，用人单位却不要了，搞得我们很被动，工作比较难做啊。

Miss 陈：这么说来，你是希望较为准确地预测下一年度公司的员工需求人数，以此为基础制订公司年度招聘计划，对吗？

小肖：是的，这个问题困扰我们好几年了，真头疼，不知道有什么好

办法。

Miss 陈：其实可以试试通过数据分析的方法来预测公司下一年的员工需求人数，再结合各单位上报的需求进行矫正，就可以得到较为准确的员工需求人数了。

小肖：我们自己来预测吗？用什么方法呢？

Miss 陈：是的，可以用回归分析的方法进行员工需求人数的预测。

小肖：好像听说过这种方法，不过没深入了解过，看来得向您请教了。

Miss 陈：这样吧，你先去找一些回归分析方法应用方面的资料，然后准备一些公司的历史数据，包括经营数据、人员数量等，年份越多越好。准备好后，我们再继续谈回归分析。给你一周的时间吧。

小肖：好的，我马上去处理。

3.2 分析方法

3.2.1 回归分析

一周后，小肖来的 Miss 陈的办公室。

Miss 陈：准备得怎么样？

小肖：经理，我查阅了一些回归分析方法应用的知识，感觉很有收获呢。

Miss 陈：那我先问个问题，请你说说什么是回归分析？

小肖：嗯，这点我专门下功夫研究了一下，基本了解回归分析的来龙去脉。首先，我很好奇回归这个词是什么意思，于是在网上查了资料，发现这个词是由英国遗传学家 Galton 首先提出的。在不太遥远的 100 多

年前,Galton发现了一种现象:他发现父亲高,往往子女也高,父亲矮,子女也矮;但是当父亲很高时,他的儿子一般不会比父亲更高,当父亲很矮时,他的儿子一般不会比父亲矮,儿子的身高会向一般人的均值靠拢。这位遗传学家将这种现象称为"向均数回归",从此产生了"回归"这样一个概念。

Miss陈:很好,了解历史是学习知识的有效方法,那你再说说什么是回归分析。

小肖:回归分析是最为常用的寻找影响因素的统计分析方法。包括两个组成部分:因变量和自变量。因变量顾名思义,就是因为某些原因而产生变化的变量,是对结果的描述,多数情况下只有一个因变量;自变量可想而知,就是自身发生变化的变量,是影响结果的各种原因的描述。自变量可以是一个,也可以是多个,通常都会有多个自变量。比如,我们等会儿要用公司的经营数据去分析和预测员工需求数量,那么自变量就是经营数据,因变量就是员工数量。

Miss陈:不错,看来你功课做得挺仔细啊,那么回归分析又有哪些类别呢?

小肖:这方面我也查了些资料,发现回归分析是个大家族,有多种类型的回归分析,最常见的有线性回归、logistic回归、cox回归,等等。

Miss陈:很好,已经很接近我们要用的分析方法了,再问一个问题,你说说什么是线性回归?

小肖:这方面不是搞得太清楚,不过我知道线性回归根据自变量的个数的不同,分为一元线性回归和多元线性回归。这里的"元"指的就是自变量的个数。比如,我们在进行经营数据和员工人数之间的回归时,如果选择"公司年度经营收入"作为自变量来进行回归分析,那么就叫作一元线性回归,因为只有一个因变量和一个自变量。如果我们同时把"公司年度经营收入""净利润"作为自变量来进行回归分析,那么就叫作多元线

性回归。不过我对“线性”了解得不多，不知道为啥叫线性。

Miss 陈：好，我接着你的内容往下说。“线”指的是坐标系中的直线，“线性”就是说自变量和因变量之间大致呈直线函数关系。注意，不是指标准的直线，而是大体呈现直线关系。举个例子吧，比如我们通信分公司营销人员的收入，其中的绩效工资主要是靠销售提成获得，假如每卖出 1 部手机，提成 100 元，那么卖得越多收入就越高，是不是？

小肖：是的，但是这和线性回归有什么关系呢？

Miss 陈：我们可以画一个坐标图，x 轴代表销售手机的数量，y 轴代表收入提成，把通信分公司营销人员的实际情况画到坐标图中，每一个点代表一个销售人员销售手机的数量，如图 3-1 所示。

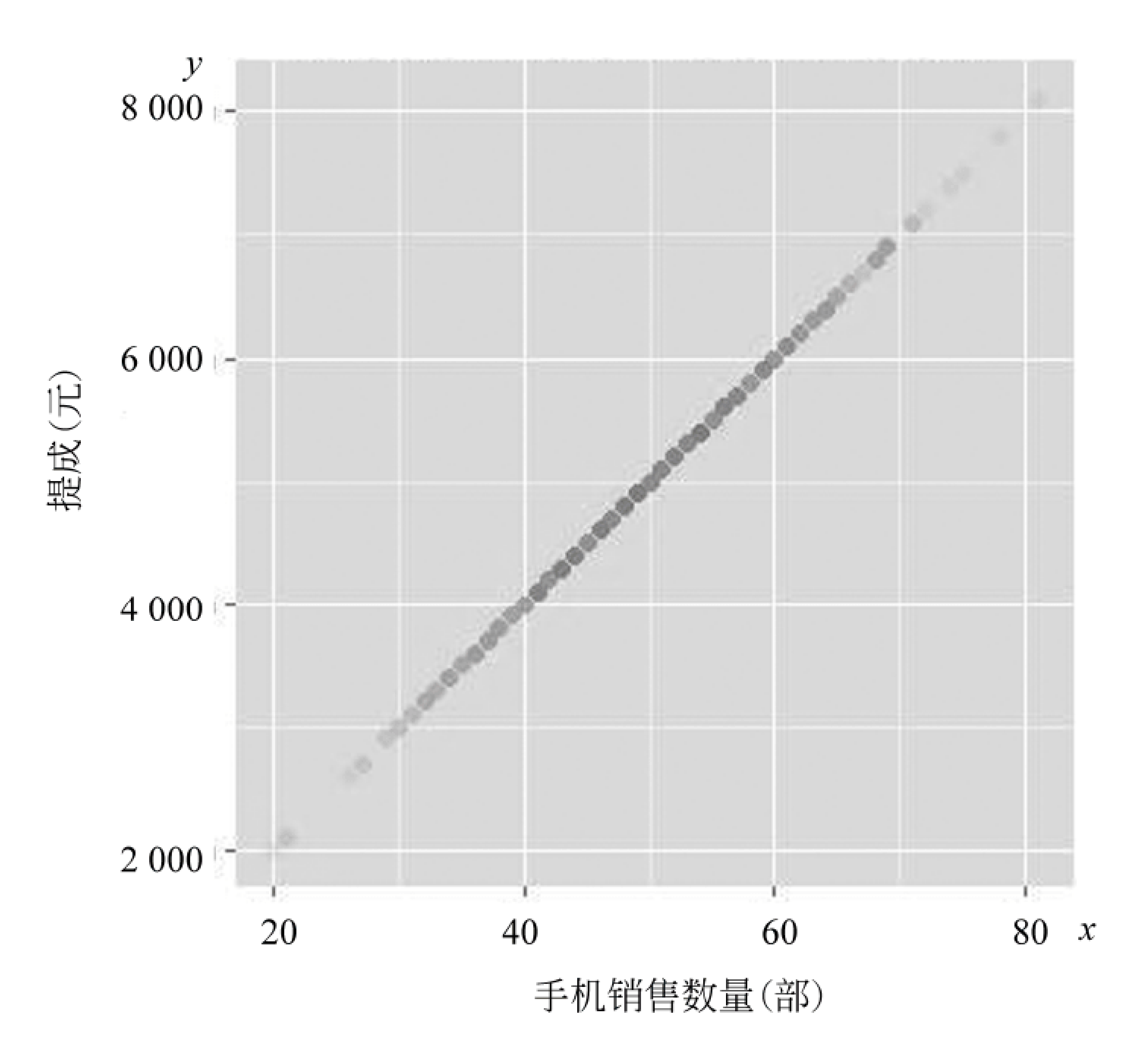


图 3-1　通信分公司销售人员销售收入与提成关系图(1)

小肖：嗯，看到了，这些点看起来好像一条直线，这就是线性关系吗？

Miss 陈：是的，如果把这些点用一条直线连起来，那么这条直线就

叫作回归线，其中手机销售数量就是自变量，提成就是因变量，它们的回归方程是：

提成=手机销售量×100

如图 3-2 所示。

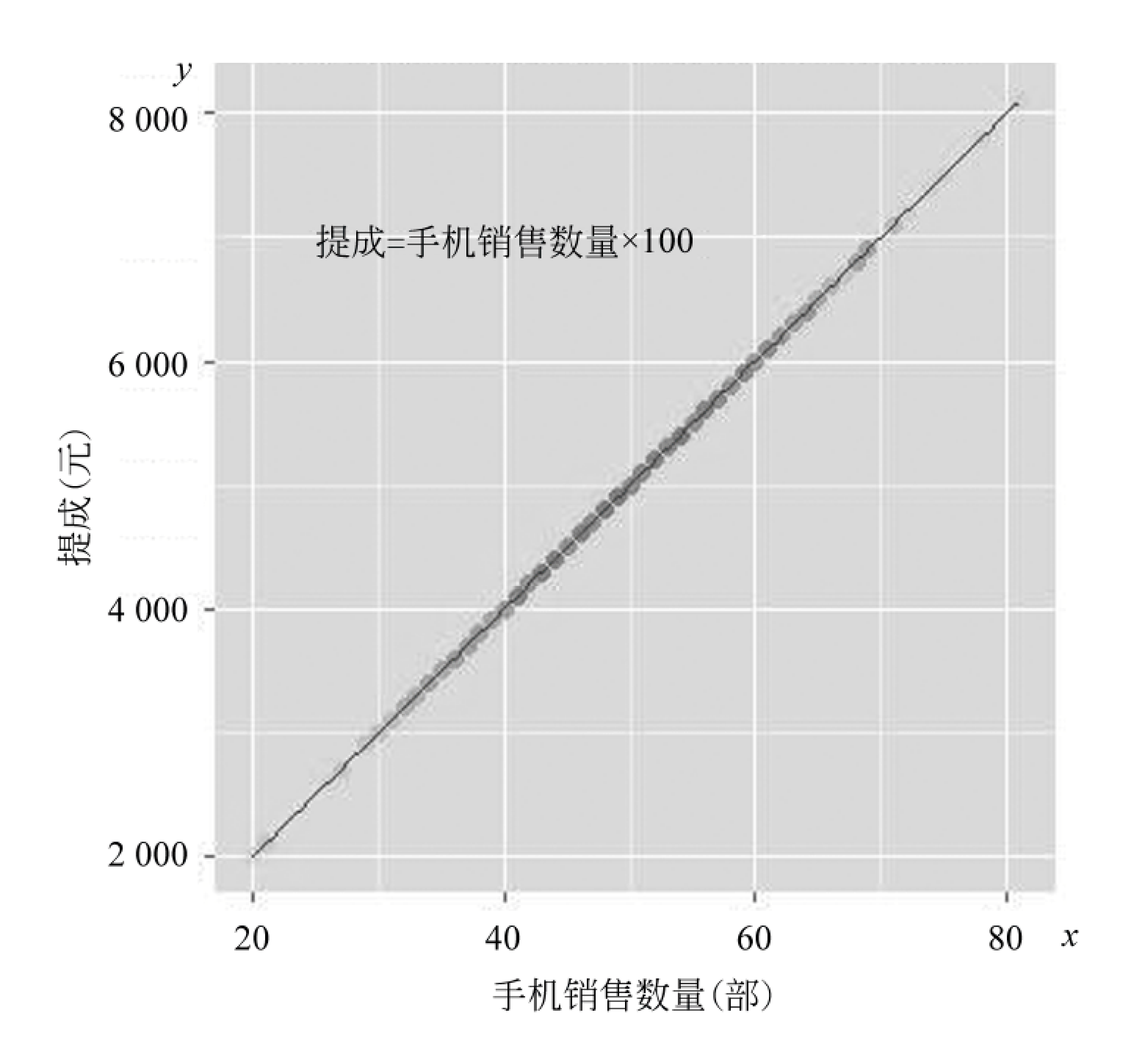


图 3-2 通信分公司销售人员销售收入与提成关系图(2)

小肖：您刚才说的回归方程是什么意思呢？

Miss 陈：回归方程就是这条直线的函数表现形式，如果是一元回归方程，那么方程式如下所示。

$$y=ax+b$$

其中，y 是因变量，x 是自变量，a 是直线的斜率，b 是直线的截距。刚才关于营销人员收入提成的例子中，y 就是提成，x 就是手机销售数量，a 就是 100，b 为 0。

小肖：哎呀，这些知识在中学学过呢，您一说就想起来了。不过，我

们的实际工作很少碰到用这些数学知识的情况。对了，多元回归方程的方程式又如何表述呢？

Miss 陈：多元回归方程的方程式如下所示。

$$y=a_1x_1+a_2x_2+\cdots a_nx_n+\varepsilon$$

其中，$a_1\sim a_n$ 叫多元回归方程的参数，ε 是误差项。

小肖：看上去多元回归的方程式有些复杂呢。

Miss 陈：自变量多了，回归方程自然会复杂些，而且方程求解的方法和过程也较复杂，不过现在有很多统计软件都可以快速求解回归方程的参数值，倒不用担心计算的复杂性问题。

回到刚才关于手机销量的例子。这个例子很特殊，因为手机销量和提成本来就是很明显的直线关系，提成就是根据销量计算出来的，它们之间是等比例关系，所以图中的回归线是一条标准的直线。但是在实际环境中，自变量和因变量很少会有这种标准的函数关系，大多数时候自变量和因变量并没有直接的线性关系，更多是一种相关关系。

比如，员工的绩效一般会受到学历水平、工作年限等因素的影响，根据我们的经验，会认为学历越高、工作经验越丰富的员工其工作绩效往往也较高，但学历、工作经验和工作绩效之间不是因果关系，所以咱们不能说学历高、经验丰富的员工的工作绩效就一定会高，不是这种关系，只是近似的推理，实际上会有偏差。只是从总体范围来看，学历高、经验丰富的员工，绩效高的可能性会更大，所以绩效高的人也会更多，我们将这种关系称为相关关系。

小肖：那这种情况能进行回归分析吗？

Miss 陈：当然可以，只要自变量和因变量之间存在相关关系，就可以尝试进行回归分析，建立回归方程。

小肖：那怎么知道自变量和因变量之间是否有相关关系，相关关系程度如何呢？

Miss 陈：判断相关关系可以计算相关系数。举个例子，比如我们某个分公司开展员工绩效考核，会得到三类分数，分别是绩效总分、适应总分和情绪总分。原始数据见表 3-1。

表 3-1　某分公司员工绩效考核结果

ID	员工编号	性别	部门	绩效总分(分)	适应总分(分)	情绪总分(分)
1	1	1	1	12.00	11.00	12.00
2	2	1	3	13.00	10.00	12.00
3	3	1	1	20.00	10.00	14.00
4	4	2	2	8.00	12.00	8.00
5	5	2	3	11.00	12.00	12.00
6	6	2	1	11.00	11.00	10.00
7	7	2	3	14.00	8.00	11.00
8	8	2	1	11.00	10.00	13.00
…	…	…	…	…	…	…

在表 3-1 的数据中，性别、部门这两个变量本来是文本类型，我们进行了编号，使其数量化，转换为无序分类数据。比如性别，用 1 代表男性，2 代表女性。

下面我们计算一下三类绩效分数之间的相关系数，并分析这三类分数之间是否存在相关关系，以及相关程度如何。结果见表 3-2。

表 3-2　绩效考核结果相关系数表

	绩效总分(分)	适应总分(分)	情绪总分(分)
绩效总分	1.00	0.47	0.54
适应总分	0.47	1.00	0.41
情绪总分	0.54	0.41	1.00

计算相关系数的 R 语句如下：

```
#计算相关系数
d<-read.csv("短期绩效.csv")
cor(d[,c(5,6,7)])
```

可以看到，绩效总分和适应总分的相关系数是 0.47，绩效总分和情绪总分的相关系数是 0.54，适应总分和情绪总分的相关系数是 0.41，三类绩效分数两两之间呈现中等程度的正相关。

该结果用相关矩阵图表示如图 3-3 所示。

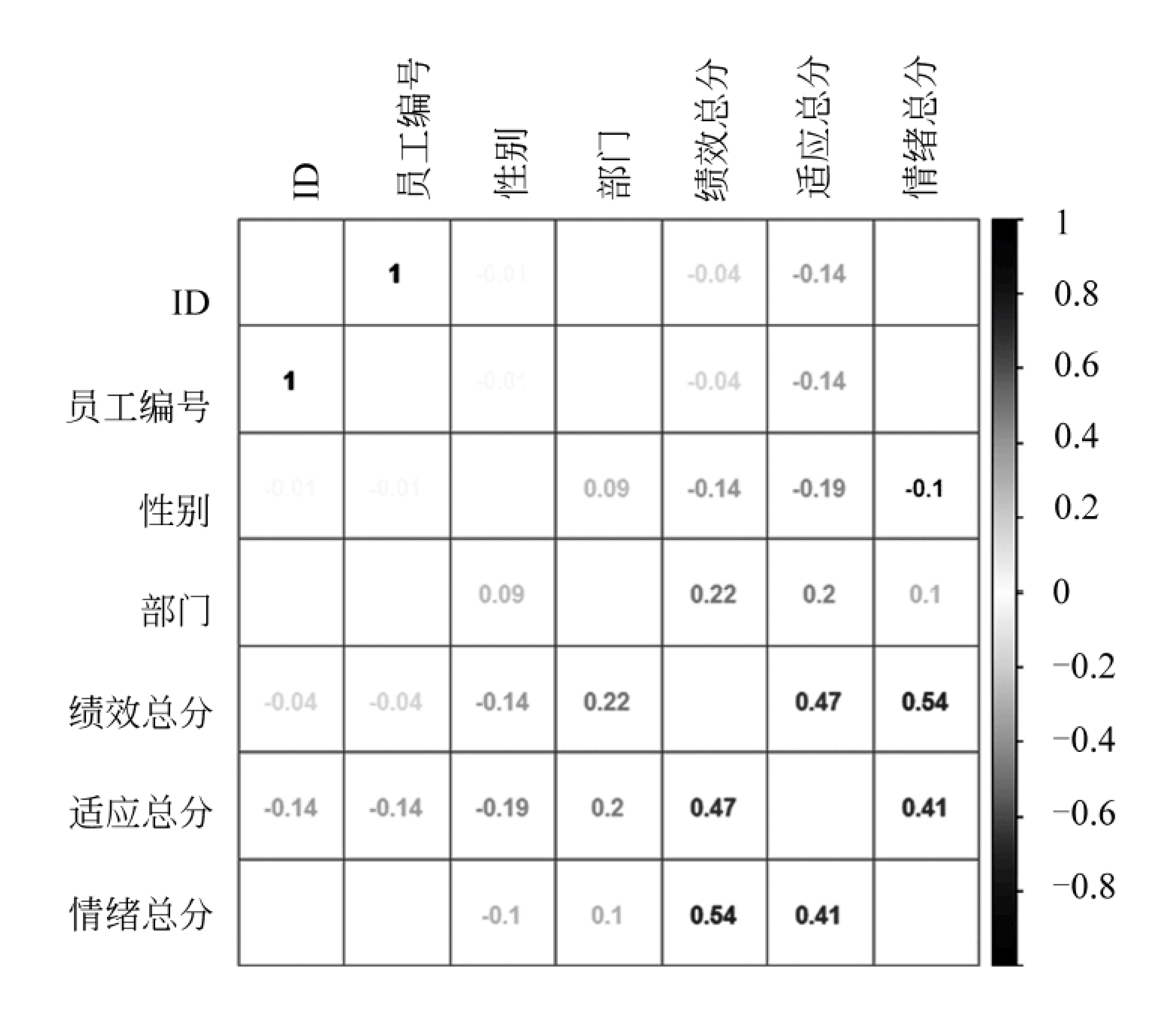


图 3-3　绩效考核结果的相关矩阵图

短期绩效的 R 语句如下：

```
library(corrplot)
d< -read.csv("短期绩效.csv")
corrplot(cor(d),method="number",diag=FALSE)
```

图 3-3 中每一个格子中的数字表示两个变量之间的相关程度，正数

表示正相关，负数表示负相关，数字的大小和颜色的深浅表示相关程度，从图中也可以看出三个绩效分数之间呈现中等程度的正相关关系。

小肖：画图的方式好像方便很多啊，变量之间的相关关系被直观地表示出来了。

Miss 陈：是的，一般都会先绘制变量之间的关系图，再初步决定用什么方法来分析。对了，说明一下，通常在进行数据分析之前，会对数据进行初步的探索，为选择合适的统计方法提供依据。这种对数据特征的探索以绘图居多，比如在分析开始前常常先画散点图、相关矩阵图、箱型图、直方图等，来研究数据的分布情况、判断数据之间的关系，等等。前面的手机销量和提成的图就是散点图。

我们来看一下绩效总分与其他两个绩效分数的散点图吧，如图 3-4 和图 3-5 所示。

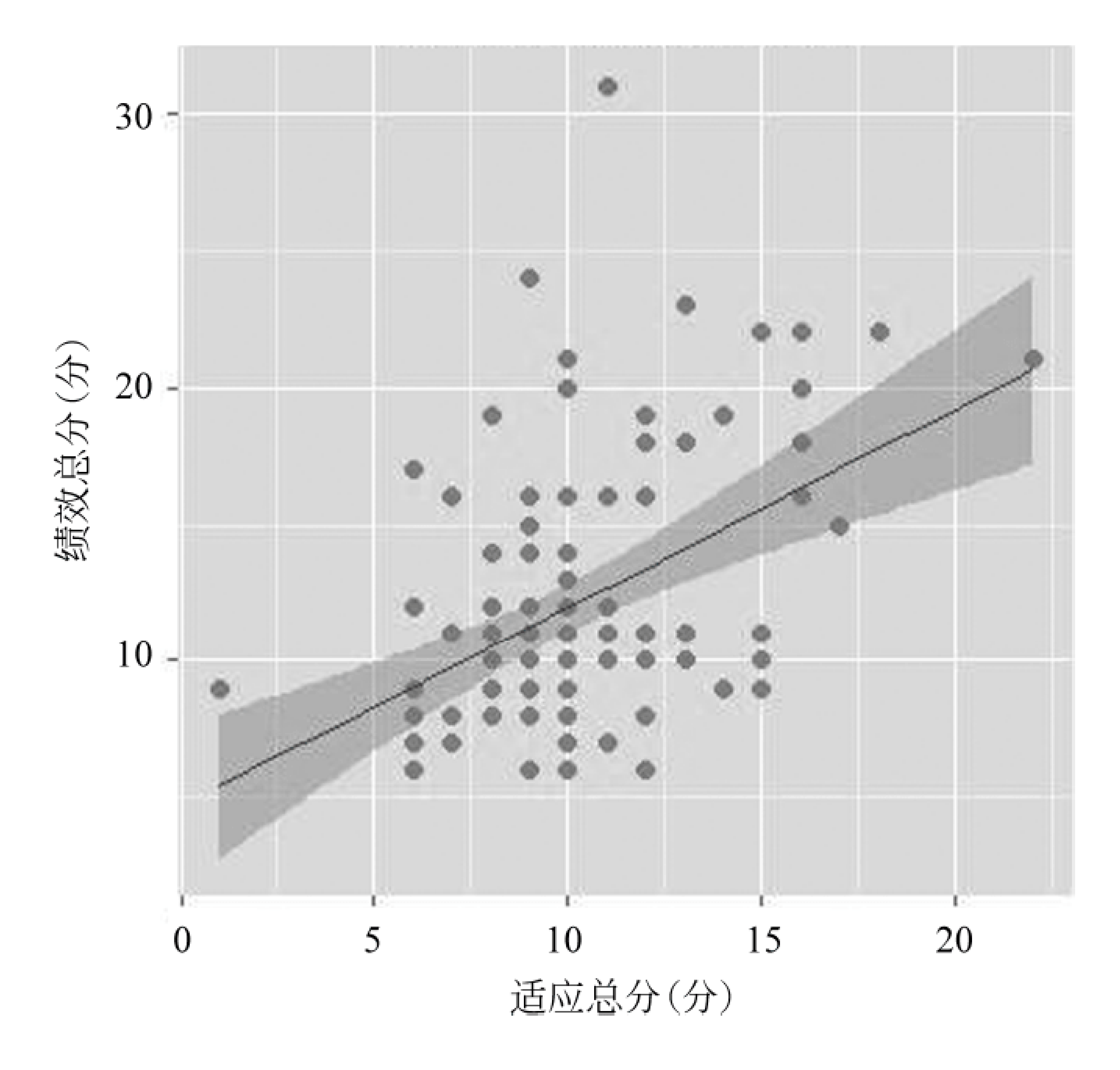


图 3-4　绩效总分与适应总分的散点图

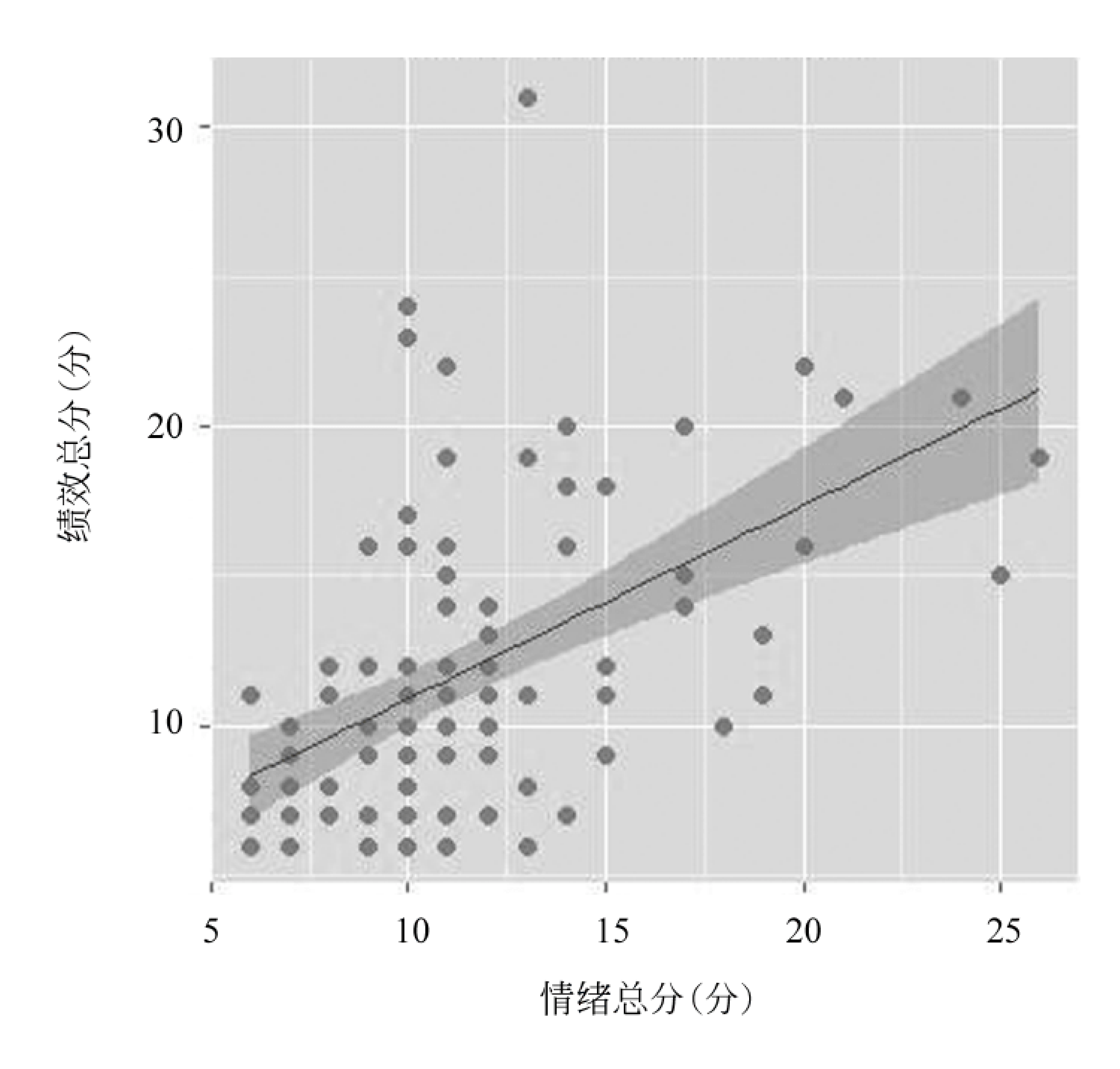


图 3-5　绩效总分与情绪总分的散点图

绘制散点图的 R 语句如下：

```
d<-read.csv("短期绩效.csv")
g<-ggplot(d)
g+geom_point(aes(适应总分,绩效总分,size=10,colour="red"))+
  theme(legend.position="none")+
  labs(title="绩效总分与适应总分的散点图")+
  stat_smooth(aes(适应总分,绩效总分),method="lm")
g+geom_point(aes(情绪总分,绩效总分,size=10,colour="red"))+
  theme(legend.position="none")+
  labs(title="绩效总分与情绪总分的散点图")+
  stat_smooth(aes(情绪总分,绩效总分),method="lm")
```

小肖：图中的直线就是回归线吗？

Miss 陈：是的，回归线周围的阴影表示因变量的置信区间。

小肖：什么是置信区间啊？

Miss 陈：简单来说，置信区间是指因变量的浮动范围，在这个范围

内因变量出现的概率为 95%(或者更高)。回归分析涉及的知识和概念比较多,我们作为企业的职能管理人员,不用钻研得太过深入,毕竟不是搞科研的,掌握基本的概念和方法,能够在实际管理中应用并给我们提供参考决策的依据就可以了。如果有兴趣,你可以看一些统计学方面的专业书籍来补补这方面的知识。

小肖:好的。经理,您刚才提到了绩效总分之间呈正相关关系,那什么是正相关关系呢,是不是还有负相关关系呢?

Miss 陈:正相关关系是指两个变量之间相关,且变化趋势相同。例如,身高和体重,一般身高越高,体重就越重,而身高越矮,体重就越轻,身高和体重的变化趋势相同,就叫作正相关关系。

小肖:这么说来,负相关关系应该是指两个变量之间相关,但是变化趋势相反。例如,随着气温升高,秋冬季节的衣服销量就会下降,但气温下降,秋冬季节的衣服销量就会上升,是这样吧?

Miss 陈:是的,你说得很正确。还有一种特殊的相关关系,叫零相关,就是说两个变量之间没有任何关系,一个变量的变化并不影响另一个变量的变化。

小肖:嗯,明白了。

3.2.2 回归分析的作用

Miss 陈:小肖,你知道回归分析有什么作用吗?

小肖:根据我查到的资料,回归分析的作用大致有两种:一是寻找事情发生的原因,比如刚才的例子,销售的手机越多,员工的提成就越高,根据分析发现手机销量和员工提成之间存在线性关系,那么手机销量就是影响员工收入的原因。二是预测,这需要用到回归方程,可以改变自变量的值来计算因变量的估计值,实现预测的目的,比如可以根据手机销量来预测员工的提成。

Miss 陈：很好，如果你发现了一种现象，又想探索这种现象背后的原因，就可以考虑采用回归分析。如果这种现象可以用连续型数值来描述，可以考虑采用线性回归。

小肖：什么是连续型数值呢？

Miss 陈：连续型数值是对数据的一种分类，像身高、体重、年龄等数据就是连续型数据，这类数据任意两点之间可以有任意个数据。比如身高，我的身高是 170cm，你的身高是 162cm，我们两个的身高之间，存在无数个值，都可以表示身高，这种数据就是连续型数值。与之相应的是离散型数值，像岗位层级、员工类别等，这类数据的任意两点之间只有有限个数据。比如，我们的员工岗位层级有 20 个级别，8 岗和 10 岗的员工之间，只存在 9 岗，不能在中间无限划分岗位层级。这类数据就是离散型数值。

小肖：那么是不是回归分析的数据一定要是连续型数值呢？

Miss 陈：原则上是的。

小肖：糟糕，我们分析人员需求时，员工人数是离散型数值啊，那不是不能进行回归分析了吗？

Miss 陈：别担心，员工人数可以看作近似连续型数值，进行回归分析。

小肖：原来是这样。

Miss 陈：其实，线性回归分析的使用条件是比较严格的，这些条件包括：①自变量和因变量之间要有线性关系；②变量要是连续型数值；③线性回归方程的残差要服从正态分布、独立性和方差齐性。其中涉及一些统计学的概念，我们后面讲到的时候再讨论。

小肖：好的。

3.3 数据准备

3.3.1 分析影响人员数量的指标并收集数据

Miss 陈：接下来要开始进行回归分析，你准备了什么数据？

小肖：我来不及找全所有公司的数据，所以先准备了两个分公司的数据，就是 A、B 两个分公司。由于我们的分公司业务类型差异太大，如果合并在一起进行人员总数的预测，会有比较大的误差，所以我打算分别进行各个分公司的人员需求预测，然后再汇总起来。

Miss 陈：很好。

小肖：我先对数据做了一些观察和分析，尝试计算了相关系数，然后根据相关系数的大小，挑选了与人员数量相关程度较大的变量来做分析。以 A 分公司和 B 分公司为例，情况如下。

A 分公司的数据包含两个变量，分别是年销售额和员工总数。因为除了年销售额这个变量之外，其他的变量和员工人数的相关系数都不大，所以就不作为分析的变量了。原始数据见表 3-3。

表 3-3 A 分公司年销售额和员工总数历年数据

年份	年销售额(万元)	员工总数(人)
2005	40 868	1 820
2006	51 357	2 150
2007	56 108	1 816
2008	86 331	2 456
2009	193 607	3 222

续表

年份	年销售额(万元)	员工总数(人)
2010	221 368	3 833
2011	278 679	4 235
2012	295 976	4 403
2013	321 555	4 832
2014	374 970	5 439

B 分公司的数据包含三个变量，分别是年出口额、年固定资产投资额和员工总数。选择年出口额和年固定资产投资额这两个变量的原因是它们与员工总数的相关程度很高。原始数据见表 3-4。

表 3-4　B 分公司年出口额、年固定资产投资额和员工总数历年数据

年份	年出口额(万元)	年固定资产投资额(万元)	员工总数(人)
2005	2 304.2	10 206.16	266
2006	6 378.87	837.12	442
2007	5 633.96	4 577.9	382
2008	5 317.78	1 465.38	436
2009	8 581.47	1 232.33	584
2010	11 725.05	7 440.174	691
2011	13 215.2	4 055.895	768
2012	17 414.86	6 295.692	954
2013	14 362.77	12 312.059	800
2014	12 627.38	12 689.914	720

3.3.2　对数据进行相关分析

Miss 陈：那么我们现在对数据进行初步分析，探索一下数据之间的关系。

先看 A 分公司,首先绘制散点图,用图形来分析数据之间的相关程度,如图 3-6 所示。

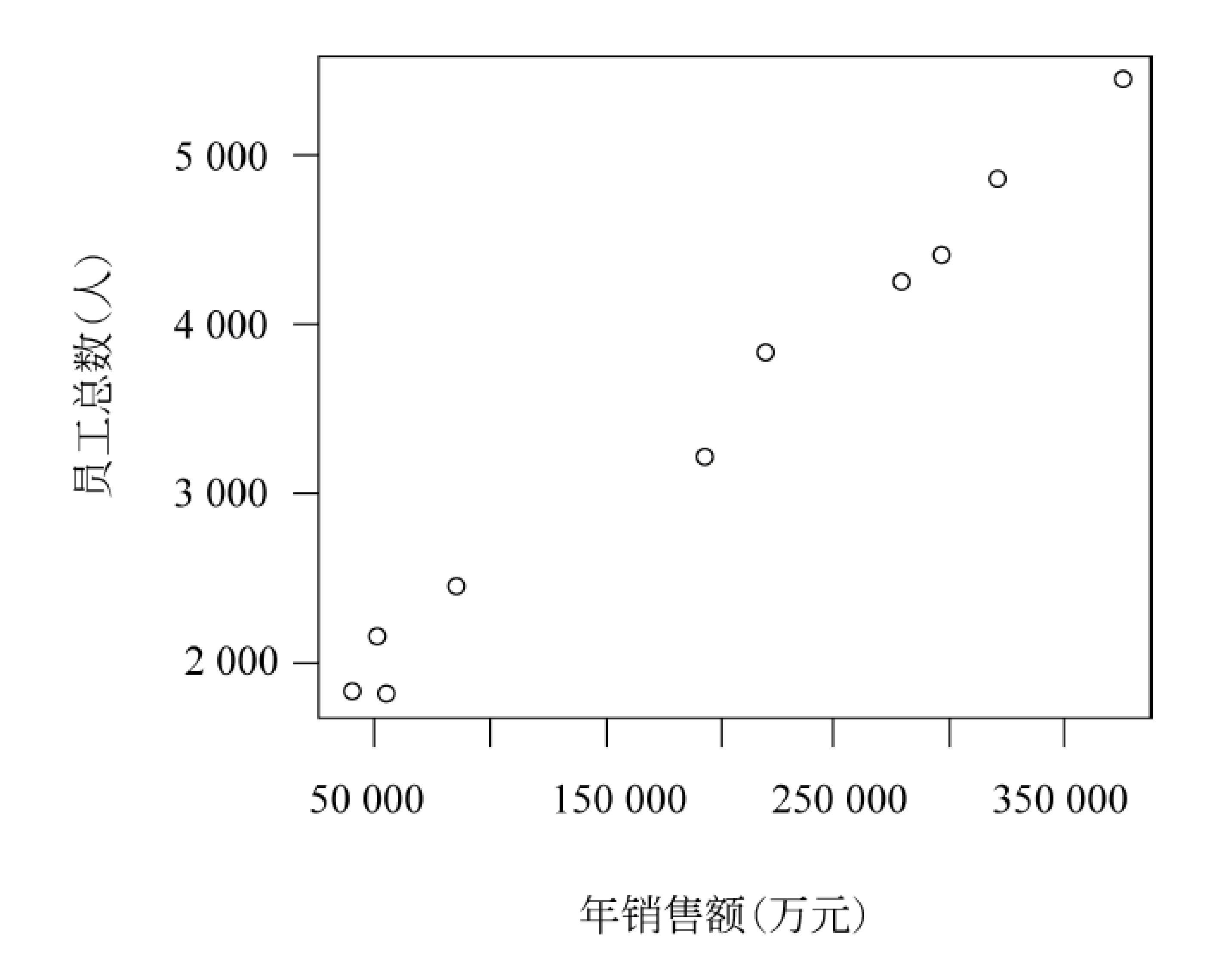


图 3-6 A 分公司年销售额与员工总数的散点图

从散点图可以看出,A 分公司的员工总数和年销售额之间存在明显的正相关关系,即年销售额越大,员工总数就越多。进一步根据其数据计算员工总数和年销售额之间的相关系数,结果是 0.99,说明两个变量之间呈现高度相关关系。

计算相关系数和绘制散点图的 R 语句如下:

```
d<-read.csv("第三章/A 分公司人员需求预测.csv")      #读取数据
cor(d[,2:3])                                        #计算相关系数
plot(d[,2:3])                                       #散点图
```

再看 B 分公司,绘制散点图,观察三个变量之间的关系,如图 3-7 所示。

从散点图可以看出,员工总数和年固定资产投资额、年出口额都存在明显的正相关关系。分别计算其相关系数,均表明三个变量之间存在高度的正相关关系,结果见表 3-5。

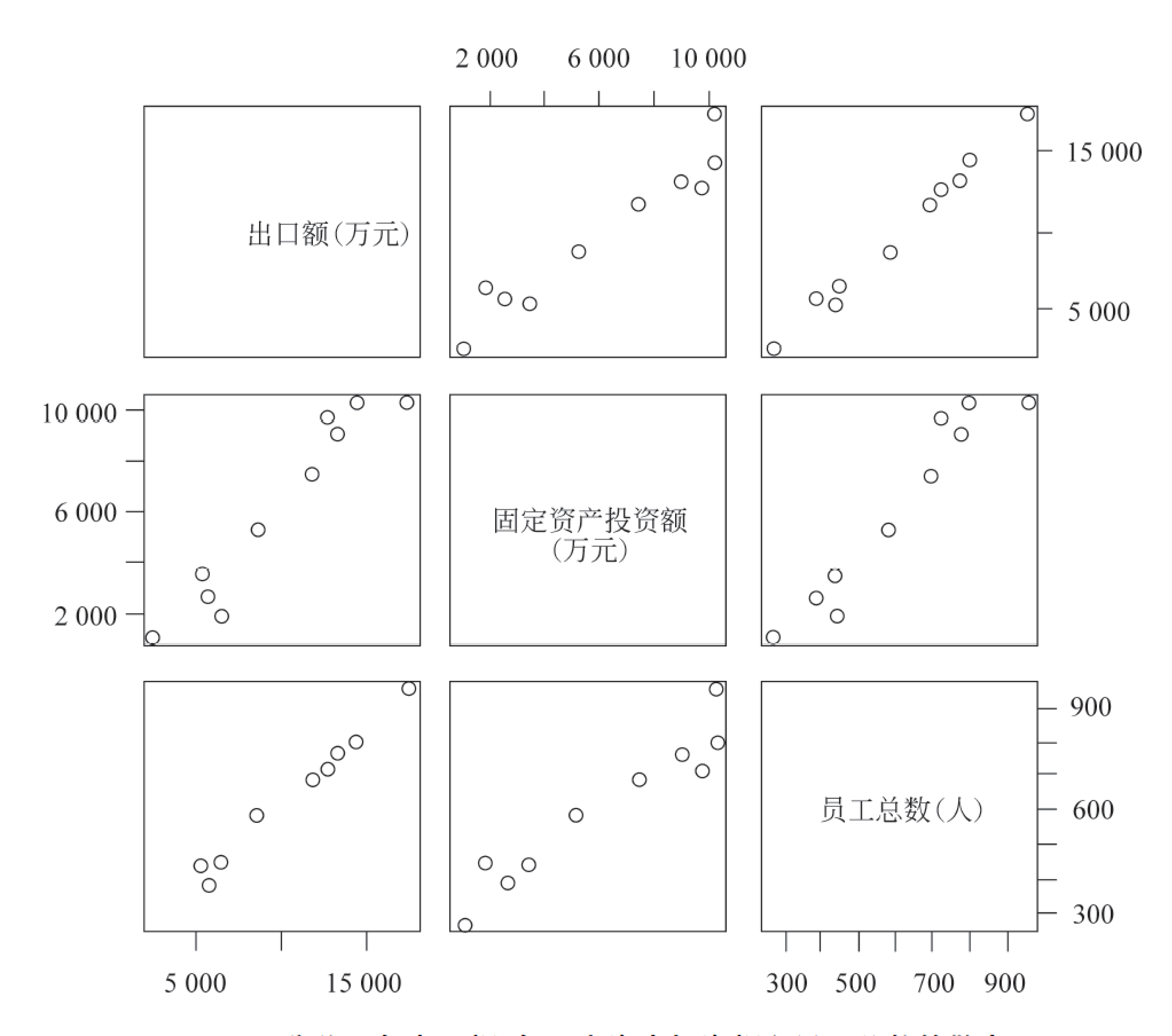


图 3-7　B 分公司年出口额、年固定资产投资额和员工总数的散点图

表 3-5　B 分公司年出口额、年固定资产投资额和员工总数的相关系数

	年出口额(万元)	年固定资产投资额(万元)	员工总数(人)
年出口额	1.00	0.96	1.00
年固定资产投资额	0.96	1.00	0.96
员工总数	1.00	0.96	1.00

计算相关系数和绘制散点图的 R 语句如下：

```
d<-read.csv("第三章/B 分公司人员需求预测.csv")       #读取数据
round(cor(d[,2:4]),digits=2)                        #计算相关系数
plot(d[,2:4],main="B 分公司员工总数散点分布图")      #散点图
```

小肖：嗯，是的，和我筛选数据变量时计算的相关系数是一样的，所以这两个分公司用了不同的变量。如果相关程度不高，那么就不适合用来进行回归分析，是这样吗？

Miss 陈：我们要进行的是回归分析，通常指的是线性回归，这种线性是变量之间存在近似直线关系，所以可以用相关系数来大致筛选维度。不过这样进行筛选仍然比较粗糙，有些维度可能和员工人数存在某种非线性关系，比如指数关系、对数关系等，这类情况就很难用相关系数来判断了。

小肖：哎呀，那不是会漏掉一些重要的维度吗？那么有什么方法可以判断维度之间的非线性关系呢？

Miss 陈：最新的一种算法叫 MINE 算法，可以探测变量之间的线性和非线性关系。如果变量之间不是相关关系，而是存在某种曲线关系，也能分析出来。但这个算法不是我们讨论的重点，以后有机会再谈吧。

3.4 分析过程：建立线性回归模型

小肖：确认了变量之间的相关关系，那么接下来该怎么分析呢？

Miss 陈：接下来我们进行回归分析。

先看 A 分公司，以“员工总数”作为因变量，“年销售额”作为自变量，进行回归分析，建立回归模型，分析结果如下：

```
Call:
lm(formula=工总数~年销售额, data=d)

Coefficients:
(Intercept)  年销售额
  1.418e+03   1.042e-02
```

根据分析结果，线性回归方程的截距(intercept)为 1 418，年销售额的系数为 0.010 42，由此可列出员工总数与年销售额的回归方程为

员工总数＝0.010 42×年销售额＋1 481

A 分公司回归分析的 R 语句如下：

```
a<-lm(员工总数~年销售额,d)          #回归分析
summary(a)                          #显示回归分析结果
coef(a)                    #显示回归方程的参数估计的结果(回归方程的系数)
```

小肖：回归分析这么快就分析完了？

Miss 陈：呵呵，是的。在 R 语言中，线性回归用 lm 函数，只需要一条语句，回归分析建模就完成了。

那么，下面看 B 分公司，以“员工总数”为因变量，“年出口额”和“年固定资产投资额”为自变量，进行回归分析，建立回归模型，分析结果如下：

```
Call:
lm(formula=员工总数~年固定资产投资额+年出口额, data=d)

Coefficients:
     (Intercept)     年固定资产投资额          年出口额
       1.678e+02         2.382e-03         4.325e-02
```

根据分析结果，线性回归方程的截距为 167.8，年固定资产投资额的系数为 0.002 382，年出口额的系数为 0.043 25，由此可列出回归方程为

员工总数＝0.002 382×年固定资产投资额
＋0.043 25×年出口额＋167.8

B 分公司回归分析的 R 语句如下：

```
a<-lm(员工总数~年固定资产投资额+年出口额,d)        #回归分析
summary(a)                                        #显示回归分析结果
coef(a)                    #显示回归方程的参数估计的结果(回归方程的系数)
```

小肖：看来进行回归分析不难嘛，很快就把模型建好了。

Miss 陈：从上面的过程来看，用 R 语言进行回归分析建模的过程的确比较简单，但是要得到一个准确的、理想的回归方程可不是一件容易的事情，还有很多工作要做，诸如以下方面。

(1) 回归模型是否在统计学上达到显著水平，也就是说模型是否有效、能用。

(2) 自变量的系数(参数估计)是否在统计学上达到显著性水平？

(3) 是否存影响作用不大的自变量？

(4) 自变量之间是否存在交互作用？

(5) 是否存在异常值？

(6) 回归方程的残差是否符合正态分布、均值为零？

上面列出的问题都是影响回归模型效果的一些因素。想要回归模型做得严谨、准确，在应用回归模型进行预测的时候能够得到精准的、符合实际情况的结果，那么就需要对上述的问题一一进行分析、验证和改进，不断进行模型优化和模型诊断。

小肖：哇，原来还要做这么多的事情啊，看来进行回归分析还挺复杂的，不像刚才想的那么简单。

Miss 陈：是的。我们看看 B 分公司回归模型的具体情况。

```
Call:
lm(formula=员工总数~年固定资产投资额+年出口额, data=d)

Residuals:
    Min       1Q    Median     3Q      Max
-35.644  -11.713  -2.794   8.099  32.554

Coefficients:
                  Estimate Std. Error t value Pr(> |t|)
(Intercept)       1.678e+02  1.955e+01   8.584   5.8e-05 ***
年固定资产投资额  2.382e-03  7.813e-03   0.305   0.769 305
年出口额          4.325e-02  5.991e-03   7.220   0.000 174 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.53 on 7 degrees of freedom
Multiple R- squared:  0.990 9,Adjusted R- squared:  0.988 3
F-statistic: 382.2 on 2 and 7 DF,  p-value: 7.118e-08
```

总体来说,B 分公司的回归方程是显著的(注意上面内容中最后一行的 p-value:7.118e −8<0.01),说明方程总体是有效的,具有统计学上的意义。再看系数,其中 Intercept、年出口额的系数都达到了很高的显著性水平,但是年固定资产投资额的系数并不显著(等于 0.769 305,远大于 0.01),说明这个自变量对因变量的影响不大,回归方程需要进行优化。

上面分析结果的 R 语句如下:

```
summary(a)          #显示回归分析结果
```

小肖:遇到这种情况我们该怎么办呢?

Miss 陈:我们需要对方程进行优化,筛选重要变量,去掉不重要变量。现在我们重新对 B 分公司的数据进行回归分析,这次在原先模型的基础上采用逐步回归方法来优化模型,结果如下:

```
Call:
lm(formula=员工总数~年出口额, data=d)

Residuals:
    Min       1Q    Median      3Q     Max
-36.771  -11.298  -2.411    7.264   32.569

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.652e+02  1.652e+01   10.00 8.47e-06 ***
年出口额     4.501e-02  1.533e-03   29.36 1.96e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
```

```
Residual standard error: 22.16 on 8 degrees of freedom
Multiple R- squared:  0.990 8,Adjusted R- squared:  0.989 7
F-statistic: 862.1 on 1 and 8 DF,  p-value: 1.962e-09
```

从上面逐步回归分析的结果可以看出，年固定资产投资额这个维度已经被自动筛掉了，筛掉之后回归方程的各项指标都达到了统计学上的显著性要求，这时的回归方程比之前的方程更为理想，回归方程也因此纠正为

$$员工总数=0.045\,01\times 年出口额+165.2$$

对回归模型进行优化，逐步回归分析的 R 语句如下：

```
a<-step(a)          #使用逐步回归优化回归方程
summary(a)          #显示回归分析结果
```

小肖：哦，原来年固定资产投资额并不是一个理想的变量啊，看来通过对回归模型的优化，可以筛选掉那些对因变量影响作用不大的自变量。

Miss 陈：是的。严格来说，接下来还要进行回归诊断，主要是对残差（预测值和实际值之间的差值）进行正态性检验、分析异常值对回归方程的影响、自变量之间是否存在多重共线性等问题。不过，我们毕竟不是搞科学研究的，对回归方程做到上述优化即可，不用再进行残差等分析，即可在人力资源管理中应用了。

3.5 结果应用：根据回归模型预测下一年度员工需求

小肖：现在回归方程已经有了，是不是就可以进行下一年度员工需求的预测了呢？

Miss 陈：是的，现在只需要把下一年度的自变量数据代入回归方程

就可以进行预测了。

小肖：嗯，我跟市场部的同事拿到了这两个分公司明年的预算数据，其中 A 分公司明年的销售额预算为 40 亿元，B 分公司的年出口额预算为 1.8 亿元。根据前面的回归方程，将数据代入方程后，计算可得两个分公司明年的员工人数为

A 分公司：5 588 人

B 分公司：975 人

太棒了，竟然这样预测出了明年需要的员工人数。

Miss 陈：不过还没结束，预测的人数实际上是有一定的浮动范围，你应该把浮动的范围也计算出来，作为预测的结果。

小肖：怎么计算预测人数的浮动范围呢？

Miss 陈：可以用 R 语言中的 predict 函数进行计算，结果见表 3-6。

表 3-6　A、B 分公司下一年度员工需求人数预测值

	预测值	最小值	最大值
A 分公司	5 588	5 173	6 004
B 分公司	975	914	1 036

根据分析结果可以知道，A 分公司明年的员工人数预计需要 5 588 人，最低需要 5 173 人，最多需要 6 004 人，实际人数落在这个范围的概率为 95%。B 分公司明年的员工人数预计需要 975 人，最低需要 914 人，最高需要 1 036 人，实际人数落在这个范围的概率为 95%。

A、B 分公司下一年度员工需求预测的 R 语句如下：

```
#A 分公司人员预测
次年销售额预算<-data.frame(年销售额=400 000)   #设置次年销售额预算值
predict(a,次年销售额预算,interval="prediction",level=0.95)
                                                  #预测(置信区间为 95%)
#B 分公司人员预测
```

```
次年经营预测数据<-data.frame(年出口额=18 000)
                                        #设置次年年出口额预算值
predict(a,次年经营预测数据,interval="prediction",level=0.95)
                                        #预测(置信区间为 95%)
```

小肖：明白了。以此类推，我就可以把每个分公司下一年的人员需求预测出来了，汇总之后就是公司总体的人员需求了。

第4章

培训师评估

导语：企业组织内部培训，在选择培训讲师时往往带有主观成分，导致出现授课效果不佳的现象，影响培训效果。本章介绍如何建立企业内部培训讲师授课评分数据库，在此基础上通过计算标准分建立常模，绘制正态分布图，用定量化的方法选择培训讲师。

4.1 需求描述

小曾：经理，这周的中层管理人员培训出了点问题。

Miss 陈：什么问题？

小曾：您知道中层管理人员的培训不好搞啊，他们参加的培训不少，对培训师的要求很高。这次的培训就有不少人向我们反映，说培训师的授课水平一般，让我们下次找好点的培训师，别浪费他们的时间。

Miss 陈：你了解具体情况吗？

小曾：我跟一些参加培训的学员做了沟通，结合培训结束后填写的培训评估表中反馈的意见，总结了一下大家反映的主要问题，有如下三点。

(1) 培训师对公司不了解，讲的都是其他行业的内容，可移植性不强。

(2) 培训师讲课的风格偏学院风，理论为主，能落地实施的内容不多。

(3) 培训师过于强势，对学员要求比较严格，把学员当作在校学生对待。

Miss 陈：这次培训在策划阶段时对培训师做过评估吗？

小曾：做过一些评估。这次的培训师是同行推荐的，我们分析了他的资料，他的知名度很高，授课经历也比较丰富，给一些知名企业讲过课，提前发来的课程提纲也比较符合我们这次的培训需求。综合来看，感觉他比较适合这次公司组织的培训，所以才请他来授课。并且根据我在培训期间的观察，这位培训师本身的知识水平是很高的，经验也比较丰富，但没想到会出问题。

Miss 陈：企业培训讲究实效，培训师要在一两天内讲授有效的知识和技能，还要控制好学员的注意力，控制授课的节奏，这对培训师的授课

技能水平要求很高，也给我们选择培训师带来了难度。准确评估培训师的技能水平确实有难度，特别是外部的培训师，有些人还在外地，很难进行直接的、面对面的接触，而且我们也较少安排试讲环节，在这种情况下想请到完全符合我们实际需求的培训师确实比较困难。

小曾：是啊，有些培训师虽然知名度高，是大教授或者名企高管，但讲课不一定精彩，甚至可能很枯燥，不受学员喜欢；有些培训师口才好，但讲课内容漂浮，很难落地，听这类培训师的课就像是听演讲，精彩有余，实用不足。如何选好培训师一直是困扰我的问题。

Miss 陈：那么你有什么想法呢？

小曾：我也没有什么好的想法，不过如果我们有一个培训师评分体系，能够对其授课水平进行量化的评估就好了。就像大众点评网对餐厅的评分，每个餐厅的口味、环境、服务都有一个评分，看到评分就能知道餐厅的基本情况，这对我们选择去什么餐厅就餐有很大帮助。类似的还有淘宝卖家的评分，豆瓣电影的评分等，都能很好地帮助我们做出合理的、准确的选择。

Miss 陈：你这个想法很好，实际上也是可以做到的。我们可以建立培训师评分体系，可以给他们打打分，对培训师进行量化评估，用标准化分数来帮助我们选择合适的培训师。

小曾：要怎么建立培训师的评分体系呢？

4.2 案例分析

4.2.1 数据准备

Miss 陈：为了建立我们企业的培训师评分系统，请先准备一些数

据吧。

小曾：需要什么数据呢？

Miss 陈：每次培训结束后，不是都要让学员填写培训评估表吗？

小曾：您是指培训评估表的评分吧，用这个评分来建立培训师的评分体系吗？

Miss 陈：是的。

小曾：哎呀，早该想到用这个，我们可是积累了好几年的数据呢。在我们的评估表中，其中一项重要的内容就是对培训师授课情况的评估，主要包括授课内容、讲授方法、进度控制、授课氛围掌控等维度。表 4-1 就是我们用的培训评估表。

表 4-1　培训评估表

培训班名称		培训时间	
学员您好： 劳驾耽误您几分钟帮助完成此份调查问卷，您的评价对于改进培训工作来说非常重要。请在空白处填上合适的分数，分数为 1～10 分，其中 10 分为最高分，1 分为最低分，并在相应的位置上填写意见。谢谢您的配合。			
课程评估			
课程准备充分，内容系统、丰富，针对性强			
你在本次培训获得的知识、技能和理念能否运用到实际工作中			
课程的内容对提升您的个人能力有帮助			
课程的内容对提升您的业务能力有帮助			
课程中安排的案例与练习及培训的形式有助于加深对课程的理解和掌握			
培训师评估			
项目			（培训师）
课程结构清晰、逻辑性强，知识量适中、重点突出			
能结合企业实际授课，案例丰富，内容深入浅出			
授课内容能反映最新的技术业务知识			

续表

语言表达清晰、幽默、流畅	
能积极调动学员学习的积极性，教学互动性强，教学进度控制良好	
授课内容能提高学员的工作绩效	
授课内容具有启迪性	
学员对培训师的满意度	
授课方式灵活、丰富，能运用各种教学道具，课堂气氛活跃	
授课准备充分，态度认真	
培训服务评估	
培训班后勤支撑情况	
培训班主任对参训学员考勤管理情况	
培训班主任能否及时处理学员反映的问题	
培训班主任工作态度，是否认真、积极和严谨	
培训班主任课前教务工作安排，是否满足教学要求	
培训环境评估	
培训课室教学设备（计算机、音响等）课前准备情况，是否满足教学要求	
培训课室教学环境搭建完善，是否符合课程内容学习要求	
后勤服务评估	
餐厅服务人员的服务态度（如无就餐无须评分）	
餐厅饭菜是否做到卫生、保温、足量（如无就餐无须评分）	
就餐方式是否便捷、有序（如无就餐无须评分）	
客房清洁卫生（如无住宿无须评分）	
客房前台服务人员的服务水平（如无住宿无须评分）	
培训服务的意见及期望	
您对培训服务的意见和建议：	

Miss 陈：我们的培训评估是在培训结束之后，由学员登录公司的培训管理系统在线填写的，所以收集数据比较方便，直接从数据库中导出来

即可，可以节省数据收集和整理的时间。

小曾：是的。在培训管理系统中，通常会用“培训师评估”的平均分来代表该培训师的总体授课效果，那么我们是否可以用这个分数来评价培训师的授课水平呢？

Miss 陈：虽然可以，但直接用这个分数来评价培训师的授课水平还不够理想。这个分数是原始分数，原始分数能够给我们提供的信息是有限的。比如，一位培训师的得分是 9 分，凭感觉我们会觉得这个得分还不错，算是较高的分数。但实际情况可能是 90%的培训师的得分都大于这个分数，这时候你再想想，9 分算高分还是低分呢？

小曾：如果 90%的培训师的得分都大于 9 分，那 9 分就不算高分了。如果看原始分数不容易判断优劣，那要怎么办呢？

Miss 陈：我们可以将原始分数转换为标准分。

小曾：什么是标准分呢？

Miss 陈：这个问题我们暂时先放一放，你先收集一下最近几年的培训师评分数据，我们再看看应该如何计算标准分。

小曾：好的。这些数据都在我们公司的培训管理系统中，马上就可以导出来。这几年公司开展了大量培训，共有 1 943 名培训师进行了授课，这些培训师包括内部培训师和外部培训师，我们对每次授课都进行了评估。如果同一名培训师讲授了多次课程，我们会取平均分。部分数据见表 4-2。

表 4-2　培训师综合评分数据

序号	姓　名	综合评分(分)
1	邝榆林	9.72
2	魏文婕	8.36
3	曾彦博	9.29

续表

序号	姓　名	综合评分(分)
4	王大勇	9.31
5	赵爱玲	8.94
6	袁海航	9.44
…	……	…
1 943	肖剑萍	9.27

4.2.2　分析案例

Miss 陈：从数据来看，我们公司的培训工作做得很到位啊，这几年竟请了这么多培训师授课。不过这里的数据只是原始数据，不能满足分析的要求，需要进行转换。现在我添加一列数据，这列数据是根据原始数据计算出的标准分，你来看看。添加数据见表 4-3。

表 4-3　培训师综合评分标准分

序号	姓　名	综合评分(分)	综合评分标准分(分)
1	邝榆林	9.72	114.76
2	魏文婕	8.36	78.19
3	曾彦博	9.29	103.20
4	王大勇	9.31	103.74
5	赵爱玲	8.94	93.79
6	袁海航	9.44	107.23
…	……	…	…
1 943	肖剑萍	9.27	102.66

小曾：咦，分数转换为 100 分上下的分数了，这就是标准分吗？转换为标准分后有什么用处呢？

Miss 陈：是的，表 4-3 中最后一列的数据就是标准分。标准分能够反映某个培训师在培训师群体中的相对位置，以此来判断培训师的授课水平。转换为标准分后，只需要知道某个培训师的原始分数，就能够判断该培训师授课水平的高低了。

现在随便在表中找一位培训师来试试看。就以姓名为“李刚”的培训师为例吧，经查询他的原始分为 9.48 分，标准分为 108 分，从标准分可以看出他的授课水平超过了 50%的培训师，经过计算可以准确知道他的授课评分超过了 79.7%的培训师，评估等级是“良好”。通过上述分析可以知道该培训师的授课水平处于中上游，其综合评分标准化示意图如图 4-1 所示。

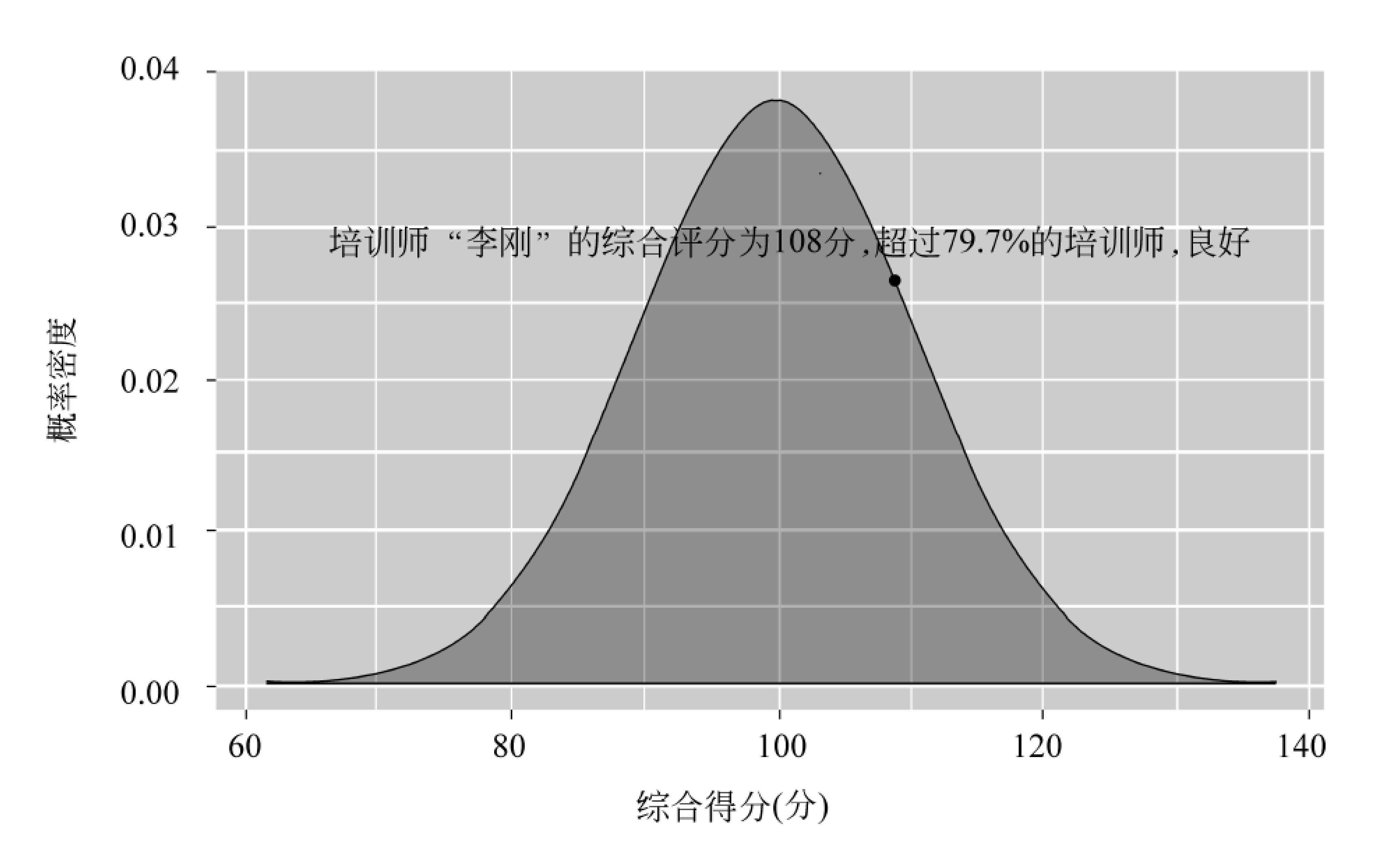


图 4-1 培训师综合评分标准化示意图

小曾：这个示意图看上去很直观啊，培训师的授课水平一目了然。

Miss 陈：把原始分转换为标准分，用标准分来评估培训师的授课水平，并区分等级的分析方法，可以比较准确地评估培训师在培训师群体中的授课水平。并且，由于分数都是我们公司的员工评出的，反映了我们公

司员工对培训师的价值倾向，所以特别符合公司的实际情况，能够最大限度地帮助我们选择符合公司实际需求，满足员工价值倾向的培训师。

不过需要注意的是，这些都是我们公司的员工评定的分数，反映的是我们公司的情况，不能推广到公司以外的地方。

小曾：哦，明白了。如果今后要聘请的培训师不在我们的数据库中，还能用这个分析方法吗？

Miss 陈：如果要聘请的培训师不在我们的数据库中，我们可以想办法将其纳入我们的评分体系。比如，我们可以找几个员工去现场听该培训师的课程，如果该培训师在网上有培训视频，还可以直接观看网上视频。然后请这些员工用相同的培训评估表对该培训师进行评分，再将评分代入上述的评分体系，计算标准分，不就把该培训师纳入我们的评分体系中了吗？

小曾：原来如此，太棒了。通过这种方式我们就可以准确评估培训师的综合水平，避免选择评分在培训师群体中位置靠后的培训师，也不用担心聘请的培训师不符合学员的需求了。不过，这种分析的过程和原理是什么呢？

Miss 陈：嗯，下面我们来看看如何进行这种分析。

4.3 分析过程

4.3.1 计算平均数和标准差

Miss 陈：你刚才看到了，我们实际上是用标准分来进行分析的。要计算标准分，就先要计算出平均数和标准差，下面就来计算培训师综合评分的平均数和标准差吧。

小曾：经理，计算平均数没问题，可标准差是什么呢？

Miss 陈：关于标准差，后面会详细讲讲。现在先简单说一下，标准差是每个数据偏离平均数的距离的平均数，是概率统计中的一个重要概念。和平均数相比，平均数反映了数据的集中程度，标准差反映了数据的分散程度，它们正好是一对。标准差的计算公式如下：

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

其中，x_i 表示每个数据，μ 表示平均数。应该说标准差是一个很常见的统计量，在各种表格、数据库、数据分析软件中都能见到，可以轻松快速地计算出来。

小曾：我试试计算一下。好了，下面是计算结果：

培训师综合评分平均数：$\mu=9.17$

培训师综合评分标准差：$\sigma=0.37$

计算平均数和标准差的 R 语句如下：

```
mean(d$综合评分)          #计算平均数
sd(d$综合评分)            #计算标准差
```

Miss 陈：做得不错！顺便提示一下，Excel 中计算平均数的函数是 average，计算标准差的函数是 stdev。计算出平均数和标准差，我们就可以计算标准分了。

4.3.2 计算标准 Z 分数和 T 分数

小曾：什么是标准分呢？

Miss 陈：标准分也叫 *Z* 分数，是通过原始分计算出来的相对位置数，反映了数据在总体中的相对位置。

小曾：那标准分怎么计算呢？

Miss 陈：标准分的计算比较简单，刚才咱们不是已经计算出了平均

数和标准差吗？标准分就是用这两个数计算出来的，公式如下：

$$z = \frac{(x - \mu)}{\sigma}$$

其中，z 表示标准分，x 表示原始分，μ 表示平均数，σ 表示标准差。

表 4-4 是根据原始分计算的标准分。

表 4-4　培训师综合评分标准分 Z 分数

序号	姓　名	综合评分(分)	综合评分标准分 Z 分数
1	邝榆林	9.72	1.48
2	魏文婕	8.36	−2.19
3	曾彦博	9.29	0.33
4	王大勇	9.31	0.37
5	赵爱玲	8.94	−0.62
6	袁海航	9.44	0.72
…	……	…	…
1943	肖剑萍	9.27	0.27

小曾：咦，这次计算出来的标准分和您刚才计算的不一样，这里的分数都在 0 上下浮动，而刚才的标准分在 100 上下浮动。

Miss 陈：很好，观察得很仔细！这里计算的是标准分 Z 分数。本来 Z 分数也可以应用，不过由于 Z 分数的量纲太小，而且还有负数，和我们习惯的百分制差别较大，所以通常又会再次进行转换，转换为标准分 T 分数。转换的方法也比较简单，公式如下：

$$T = 10 \times Z + 100$$

转换后的 T 分数变为了服从标准差为 10，平均数为 100 的正态分布数据，经过转换后的 T 分数见表 4-5。

表 4-5 培训师综合评分标准分 T 分数

序号	姓　名	综合评分	综合评分标准分 Z 分数	综合评分标准分 T 分数
1	邝榆林	9.72	1.48	114.76
2	魏文婕	8.36	−2.19	78.19
3	曾彦博	9.29	0.33	103.20
4	王大勇	9.31	0.37	103.74
5	赵爱玲	8.94	−0.62	93.79
6	袁海航	9.44	0.72	107.23
…	……	…	…	…
1 943	肖剑萍	9.27	0.27	102.66

小曾：原来如此，Z 分数转换成了 T 分数！果然，转换之后看上去就有熟悉感了，感觉像是我们的考试分数，呵呵。

4.3.3 绘制正态分布图

Miss 陈：转换为标准分后，我们就可以根据标准分计算出某个培训师在培训师群体中的位置了。

小曾：看上去，我们好像是用比例来代表相对位置的吧？

Miss 陈：是的，我们用某个培训师的标准分对应的累计概率分布值来标示其在培训师群体中所处的位置。为了正确显示培训师在培训师群体中的位置，需要先绘制一张正态分布图，如图 4-2 所示。

此图绘制过程有以下两个步骤。

(1) 随机生成若干服从正态分布的数据，这些数据服从以 100 为平均数，10 为标准差的正态分布(与培训师综合评分标准分 T 分数的平均数和标准差一致)。

(2) 根据以上数据绘制密度曲线图。

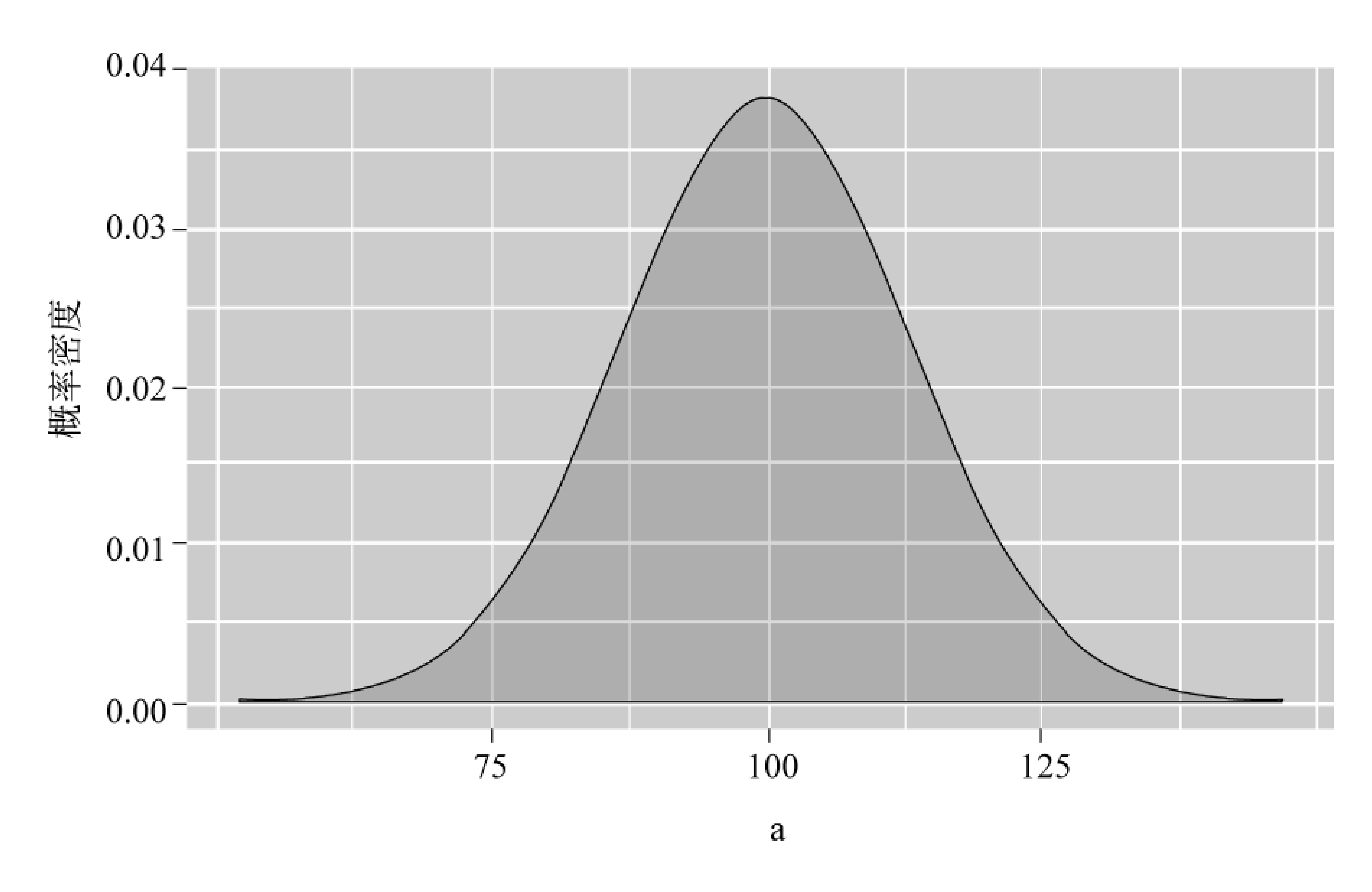


图 4-2　正态分布图

绘制正态分布图的 R 语句如下：

```
r<-data.frame(a=rnorm(10 000,mean=100,sd=10))  #随机生成 10 000
个平均数为 100,标准差为 10 的数据
g<-ggplot(r)
g+geom_density(aes(x=a),fill="blue",alpha=0.3,adjust=2)
```

4.3.4　标注位置

小曾：接下来要在图上标记培训师的位置吗？

Miss 陈：是的，标注培训师的位置也有两个步骤。

(1) 计算培训师的标准 T 分数、概率密度值、累计分布值、对应等级等数据。

(2) 根据以上数据在图中标注培训师的位置。

比如这位叫李刚的培训师，经过计算和查询，其基本情况是标准分为 108.31 分，概率密度值为 0.028，累计概率分布值为 79.7%，评价等级为

"良好"。

小曾：评定等级是怎么划分的呢？

Miss 陈：评定等级是按照标准差的大小来划分的，具体见表 4-6。

表 4-6 评定等级划分标准

等 级	判断标准	T 分数
非常优秀	≥2 个标准差	$T \geqslant 120$
优秀	1～2 个标准差	$120 > T \geqslant 110$
良好	0～1 个标准差	$110 > T \geqslant 100$
一般	－1～0 个标准差	$100 > T \geqslant 90$
较差	－2～－1 个标准差	$90 > T \geqslant 80$
很差	<－2 个标准差	$80 > T$

等级划分是人为划定的，可以根据实际情况进行调整。

小曾：原来如此。

Miss 陈：现在万事俱备只欠东风了，我们把该培训师的统计数据标示到图上去吧，如图 4-3 所示。

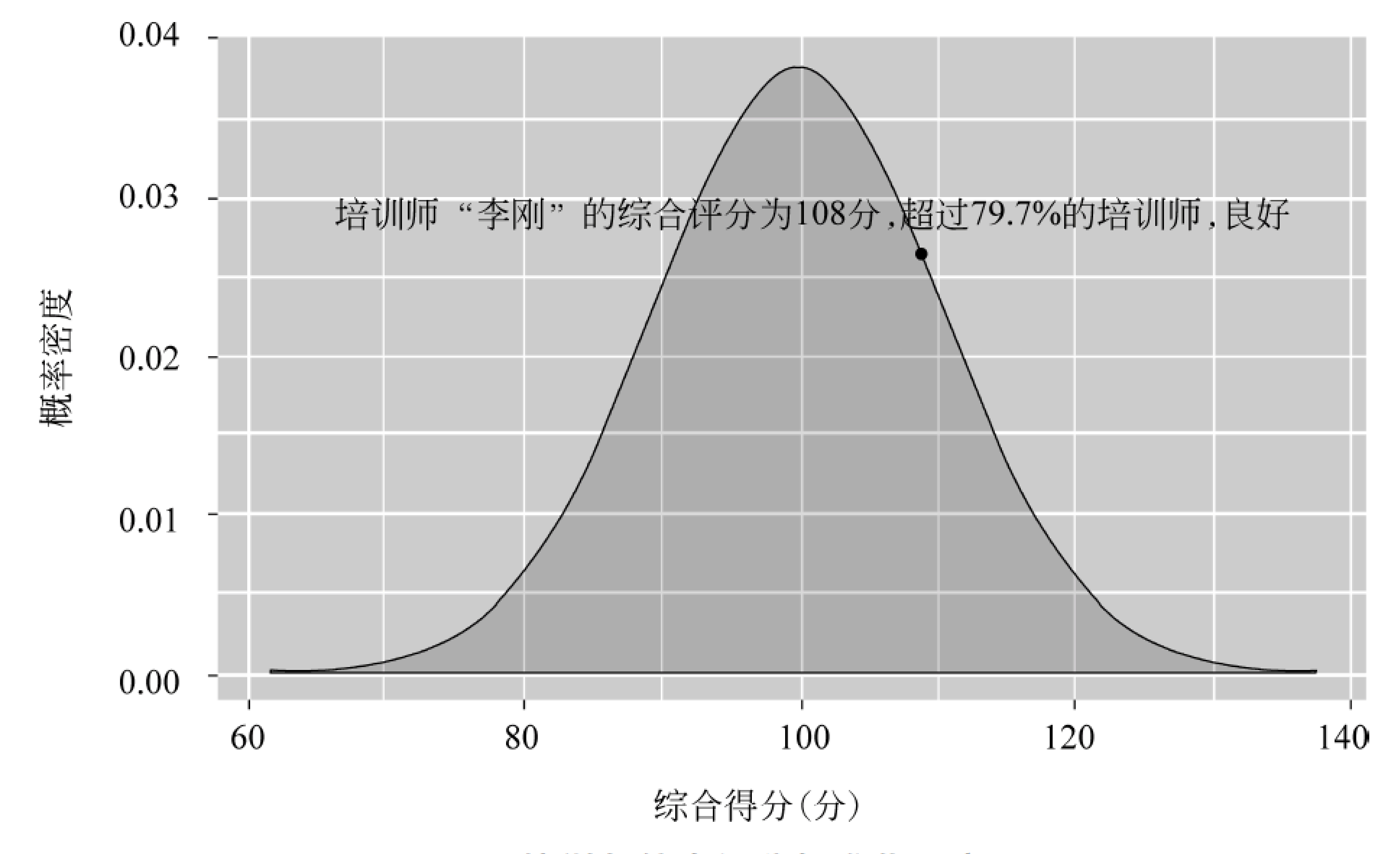


图 4-3 培训师综合评分标准化示意图

从图 4-3 可以直观看出，培训师“李刚”的授课综合评分超过了 79.7%的培训师，处于中间偏右的位置，评价等级为“良好”。

以上分析过程的 R 语句如下：

```
library(ggplot2)
d<-read.csv("第四章/培训师评分原始数据.csv")   #读取数据
mean(d$综合评分)                              #计算平均数
sd(d$综合评分)                                #计算标准差
d$综合评分标准分<-scale(d$综合评分) #计算标准分
d$综合评分标准分<-round(d$综合评分标准分 * 10+100,2) #转换为 T 分数
n<-"李刚"
x<-d[d$姓名==n,]$综合评分标准分      #提取某培训师的综合评分标准分
y<-dnorm(x,mean=100,sd=10)          #概率密度值
y1<-pnorm(x,mean=100,sd=10)         #累计分布概率
y2<-""
{
if(x< 80){y2="很差"}
else if(x<90){y2="较差"}
else if(x<100){y2="一般"}
else if(x<110){y2="良好"}
else if(x<120){y2="优秀"}
else {y2="非常优秀"}
}
l<-paste("培训师[",n,"]的综合评分为",round(x,0),"分，超过",round
(y1* 100,2),"%的人，",y2)           #生成标注
r<-data.frame(a=rnorm(10 000,mean=100,sd=10)) #随机生成正态分布值
g<-ggplot(r)
g+geom_density(aes(x=a),fill="blue",alpha=0.3,adjust=2)+
                                              #绘制密度曲线
  geom_point(aes(x,y),size=5,color="red")+    #绘制位置点
  geom_text(aes(x,y,vjust=- 1),label=l,color="red")+    #绘制标注
  labs(title="谦多顺公司培训师综合评估体系",x="综合得分",y="")
```

小曾：如果要查询和计算其他培训师的数据，是不是把上述 R 语句中变量 n 的值更改成其他培训师的名字就可以了？

Miss 陈：是的。

4.4 衍生内容

4.4.1 平均数和标准差

小曾：经理，关于标准差您能说得再详细一点吗？

Miss 陈：好的。标准差是一个很有意思的数据，和平均数有关系。要理解标准差，我们需要更进一步理解平均数。平均数是众所周知的统计量，那么我问你，平均数有什么缺点？

小曾：经常用平均数，倒没怎么想过它的缺点。不过，我们在算平均工资的时候，经常出现平均工资高过大部分员工的情况，员工都抱怨说被平均了，这算不算缺点呢？

Miss 陈：你说得很好！平均数的缺点就是容易受极端值影响。如果数据中有一些非常大或者非常小的值，平均数就会向这些数值靠拢，导致我们对数据的总体情况出现误判。

比如，我们常常看到官方公布某城市职工的平均工资，每次公布后网上都有很多人觉得平均工资太高，说自己拖了国家的后腿。某种程度上看，很可能是平均工资受到了极端值的影响，即受到那些少数的高收入人群的影响，平均数被拉高了。

小曾：明白了，由于部分人工资很高，把平均工资给抬高了。

Miss 陈：所以在某些情况下我们会用中位数来代替平均数，因为中位数更能反映实际情况。

小曾：什么是中位数呢？

Miss 陈：中位数就是把数据从小到大进行排序，处在中间位置的那个数。不过中位数和我们这次的内容相关度不大，咱还是继续说平均数

吧。你知道平均数反映了数据的什么特征吗？

小曾：您刚才说了，平均数反映了数据的集中趋势。

Miss 陈：是的，平均数反映了数据的集中趋势，也就是说反映了数据密集、集中的特性，反映了数据向中间值靠拢的趋势特征。

小曾：平均数是有这个特点。

Miss 陈：与集中趋势相反，数据还有一种特征叫离散趋势，也就是数据的分散程度。如果我们要了解数据的分散程度，就需要用到标准差。

小曾：平常很少听到离散趋势这个说法，能具体讲讲吗？

Miss 陈：举个例子吧，有 A 和 B 两组数据，如下：

A：1 2 5 8 9

B：3 4 5 6 7

你计算一下这两组数据的平均数。

小曾：好的。啊！计算出来 A、B 两组数据的平均数都是 5。

Miss 陈：平均数都是 5，那么用平均数就不能比较这两组数据之间的差异了，得想其他办法。现在我们把这两组数据投射到坐标轴上看看，如下所示。

0	1	2	3	4	5	6	7	8	9	10
	A	A			A			A	A	
			B	B	B	B	B			

小曾：看到了，A 组数据更分散，B 组数据更集中。

Miss 陈：是的，虽然这两组数据的平均数相同，但是它们的分散程度却不同，这种分散程度就叫作离散趋势，而衡量这种离散趋势的指标就是标准差。还记得标准差的计算公式吗？

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

式中，σ 为标准差；n 为数据个数；x_i 为第 i 个数据；μ 为平均数。

小曾：嗯，这公式看上去其实还有点儿复杂。

Miss 陈：其实理解起来并不复杂，标准差计算公式的意思是：每个数与平均数的差的平方和，再除以数据个数后开方。简单来说就是计算每个数与平均数的差异之和。当然，如果手动计算会比较麻烦，好在有很多软件都可以方便地计算出标准差，比如 Excel，用函数 Stdeva 就可以轻松地计算出标准差。

小曾：我来计算一下 A、B 两组数据的标准差。

A 组数据标准差：$\sigma_A=3.54$

B 组数据标准差：$\sigma_B=1.58$

数据更分散的 A 组，其标准差更大，B 组的标准差更小。这么看来，如果标准差越大，那么数据的离散程度就越大，是这样吗？

Miss 陈：是的。

4.4.2 正态分布

小曾：经理，您能讲讲正态分布吗？

Miss 陈：好的。正态分布（normal distribution）又名高斯分布（Gaussian distribution），据说是高斯先生最先应用的。对了，就是那个著名的数学家高斯。现在德国的 10 马克钱币上还印着他的头像和正态分布的密度曲线呢，以纪念这位伟大学者。

正态分布是连续随机变量概率分布的一种频率分布形式。举个例子吧，人的身高是不同的，有的人个子高，有的人个子矮，高矮胖瘦各不相同，是吧？但是，特别高和特别矮的人并不多，大多数人都是中等身高。如果把全世界的人的身高放到一起，按照身高出现的频率绘制坐标图，用横坐标表示身高，纵坐标表示人数，那么一定会得到如图 4-4 所示的频数图（直方图）。

大部分人的身高会集中在中等高度的附近，越往极端方向延伸（很高

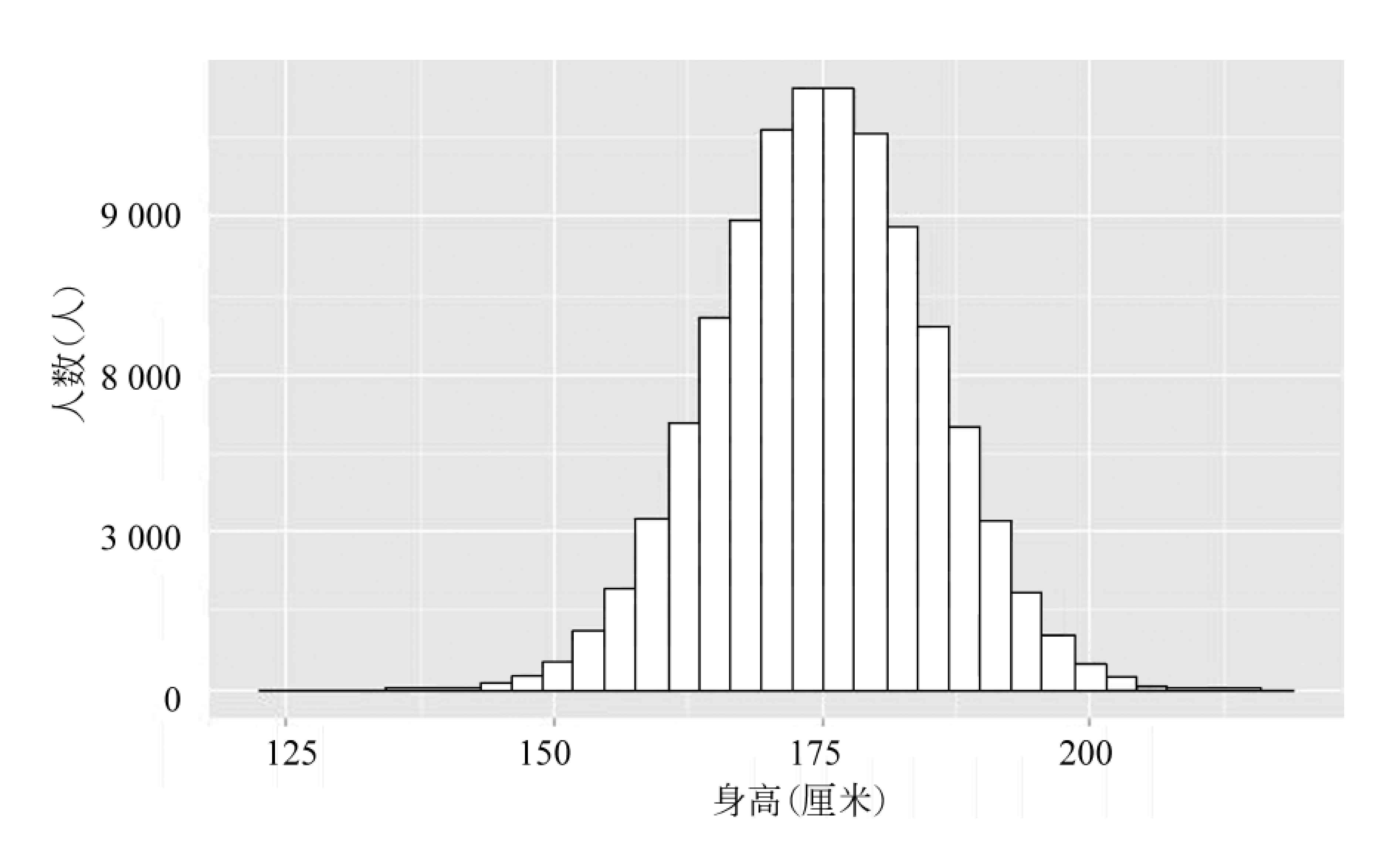


图 4-4　人体身高分布频数图

或很矮的方向)，人数就越少，这种分布就是正态分布。如果把直方图转换为密度曲线图，用概率来代替人数，就变成如图 4-5 所示的正态分布图了。

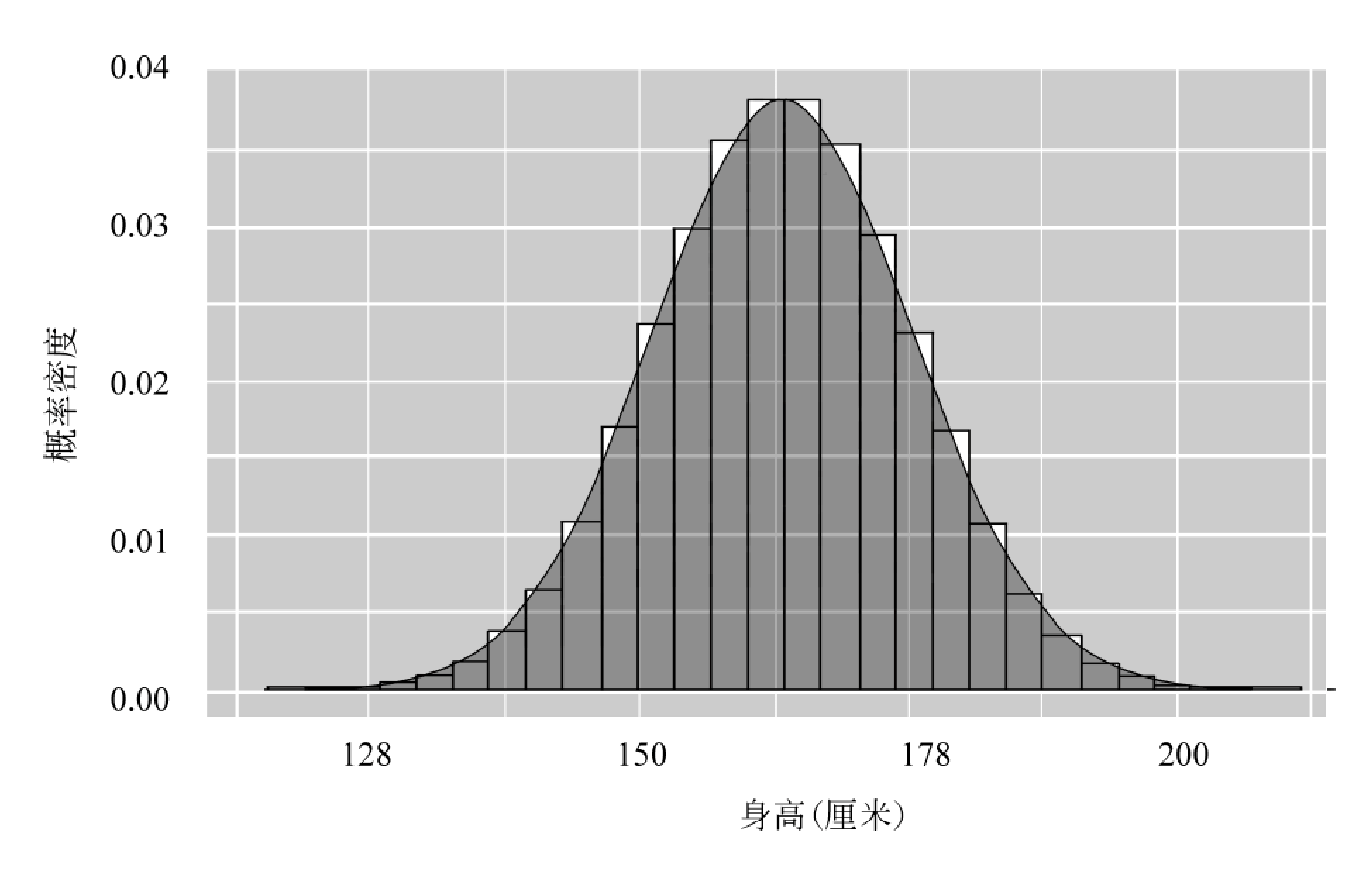


图 4-5　人体身高正态分布图

正态分布的特点有以下几方面。

(1) 正态分布是左右对称的，对称轴是经过平均数点的垂直线。

(2) 正态分布的中央点最高，然后逐渐向两侧下降，曲线的形式是先向内弯，再向外弯。

(3) 正态曲线下的面积为1。正态分布是一簇分布，受到平均数、标准差的大小与单位不同而有不同的分布形态。标准正态分布是正态分布的一种，其平均数和标准差都是固定的，平均数为0，标准差为1。

(4) 正态分布曲线下标准差与概率面积有固定数量关系。所有的正态分布都可以通过Z分数公式转换成标准正态分布。

在自然界和人类社会中存在大量的正态分布形态，比如鹅卵石的长度、高考的成绩、人的身高和体重、每年的降雨量、植物叶片的直径大小等，基本上都服从正态分布规律。

小曾：听了这些，我对正态分布的认识清晰了很多呢。但比较疑惑的是，为什么自然界和人类社会中会出现正态分布的现象呢？

Miss 陈：问得很好。自然界和人类社会为什么会出现正态分布的现象呢？据我所知，虽然很多人在使用正态分布，但是并不知道为什么会出现正态分布，为什么要用正态分布。要解释这个问题，我们需要了解统计学的历史。说来话长，你知道以下几点即可。

(1) 高斯发现了随机误差的分布服从正态分布规律。1801年，高斯将正态分布应用到天文学研究，用最小二乘法神奇地预测了谷神星的位置，并证明了随机误差的分布服从正态分布规律，这是正态分布在世界上的第一次应用。

(2) 自然界和生产中大量存在正态分布现象。1809年，法国著名的天文学家和数学家拉普拉斯发现高斯的研究后，马上将正态分布与他的中心极限定理结合起来，证明在自然界与社会生产中，一些现象受到许多

相互独立的随机因素的影响，如果每个因素所产生的影响都很微小时，总的影响可以看作是服从正态分布的。之后中心极限定理得到进一步发展，该定理发现无论总体数据呈现什么分布，只要取出的样本量足够大，都有正态分布的形式。中心极限定理和正态分布的结合为正态分布的应用奠定了基础。

(3) 正态分布在各个学科都得到证实并应用。1831 年，比利时统计学家、数学家和天文学家，被誉为近代统计学之父的凯特勒将正态分布的概念引入人口学，从此正态分布遍地开花，攻陷人口、政治、农业、工业、商业、犯罪等社会领域，并进一步攻占天文学、数学、物理学、生物学、社会统计学及气象学等自然科学领域。

从正态分布被发现、论证、应用的历史过程可以看出，正态分布是由统计学家、数学家、天文学家发现的一种自然现象，就像牛顿发现万有引力一样，正态分布也是一种自然现象。

小曾：原来正态分布是这样被发现和应用的，正态分布是一种自然现象，挺有趣的。回头我得找找更详细的资料，深入学习正态分布的知识。

4.4.3　标准分

小曾：那么为什么可以根据标准分计算相对位置呢？

Miss 陈：因为标准分服从正态分布。根据前面所说的正态分布的特点，只要知道了标准分的值，就可以计算其相对位置，计算出累计分布概率值。标准正态曲线的面积分布如图 4-6 所示。

还记得标准正态分布的特点吗？

小曾：记得，标准正态分布的平均数是 0，标准差是 1，把数据转换为标准 z 分数后就服从标准正态分布。

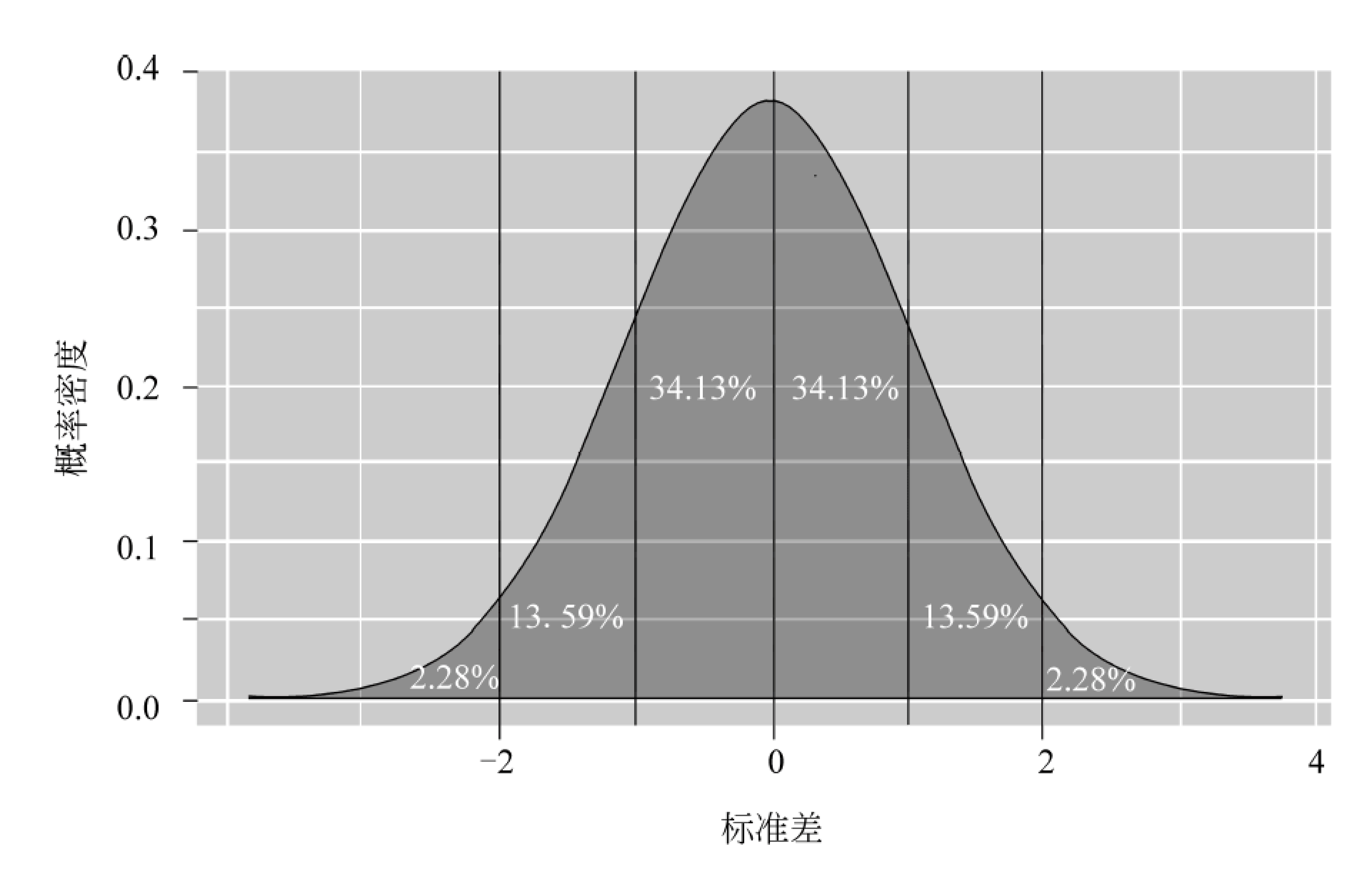


图 4-6　标准正态曲线的面积分布

Miss 陈：是的。对于标准正态分布，可以根据标准分的值计算对应的概率分布值。正负一个标准差内的面积为 34.13％×2＝68.26％，正负两个标准差内的面积为 34.13％×2＋13.59％×2＝95.44％。

以刚才的培训师评分为例，根据培训师评分可计算对应的分布面积。我们想了解培训师优于多少人，所以计算的是正态分布曲线中的左侧面积，即累计分布概率值。取培训师分数为 90 分、100 分、110 分，其对应的左侧面积如图 4-7 所示。

（本章源代码提供了网页版的正态分布演示和计算程序，该程序可以设置平均数和标准差，然后计算左、右侧面积、中间面积和双侧面积，还可以设置面积来倒推标准差，供读者练习和应用。）

小曾：累计分布概率值要如何计算呢？

Miss 陈：计算机普及前，要计算累计分布概率值，需要把数据转换为标准正态分布，然后查询正态分布表来获得分布概率值。随着计算机

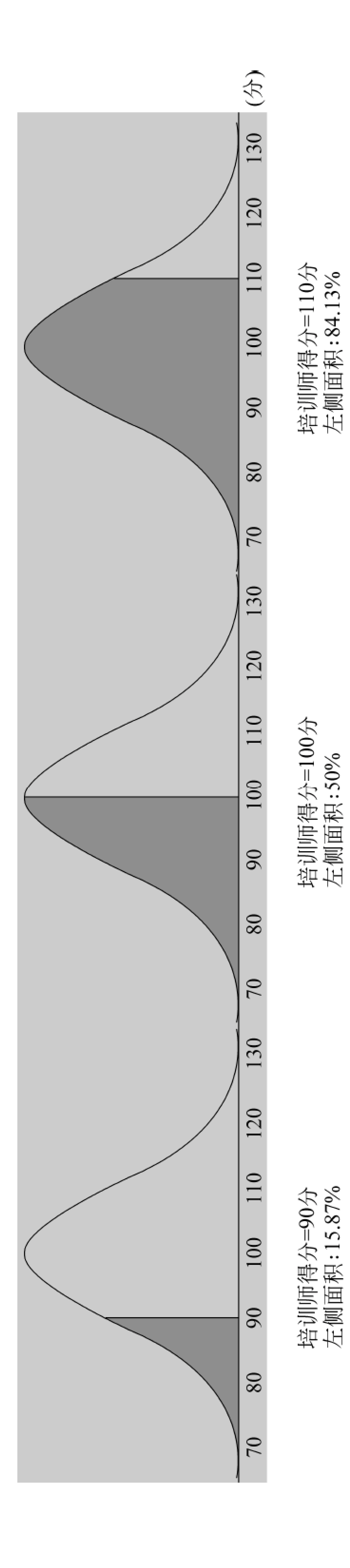


图 4-7　培训师得分与对应左侧面积图

的普及和技术的发展,以及各类分析软件的升级,大部分计算都可以通过计算机来完成,而不用像以前那样手动计算和查找。

计算机的发展让统计计算的速度得到了前所未有的提升,而且由于不用担心计算的复杂程度,近代还发展出一些更高级的统计算法。这类算法用人工计算是很困难的,比如现代流行的机器学习算法,如果不依靠计算机,将会非常的费时费力。虽然对这些算法的学习和理解有一定难度,但是依靠计算机,在实际应用的时候用一两个函数就可以完成计算,简单得多。

现在已经不需要查正态分布表,也不需要转换为标准正态分布,只需要知道平均数和标准差,代入函数就可以计算出结果。比如 Excel 中函数 NORM. DIST 就可以直接计算累计概率值,在 R 语言中用 pnorm 函数来计算。

小曾:还好计算起来比较简单,这下放心了。不过,我对 T 分数还不是太清楚,Z 分数转换为 T 分数有什么特殊意义吗?

Miss 陈:看看 T 分数的转换公式吧。

$$T = 10 \times Z + 100$$

小曾:看上去 T 分数是用 Z 分数乘 10 再加 100,这有什么含义吗?

Miss 陈:我们把上面的公式换个形式如下:

$$T = \mathrm{a} \times Z + \mathrm{b}$$

实际上,转换后的 T 分数也服从正态分布,其标准差等于 a,平均数等于 b。

所以,我们的培训师评分数据转换为 T 分数后,服从标准差为 10,平均数为 100 的正态分布。

小曾:原来是这样啊,a 和 b 就是转换为 T 分数后的标准差和平均数。

Miss 陈:是的。其实 T 分数的应用领域是挺广泛的。比如高考的

标准分，就是以 500 为平均数，100 为标准差的 T 分数；还有我们的智商测试，一般用韦氏智力量表测试的智商，都是以 100 为平均数，15 为标准差的 T 分数。如果对正态分布的分布特征比较熟悉的话，根据分数值就可以判断其大概的累积分布概率。

小曾：难怪 T 分数看着眼熟，原来高考也用了它啊。想当年对我的高考分数不甚了解，现在终于知道了。回头我得用当年的高考分数去算算我在高考大军中的位置。

第 5 章

薪酬公平性分析

导语：企业薪酬体系出现问题时，往往会使用薪酬满意度调查法来进行分析。但这类调查比较敏感，耗时较长，且效果不佳。本章介绍如何利用现成的薪酬数据，通过薪资结构图、基尼系数、薪资均衡指标、公平感计量模型等指标和方法，分析企业薪酬体系的合理性与公平性。

5.1 需求描述

小姚：经理，最近跟一些员工聊天，发现他们对薪酬有些意见呢。

Miss 陈：怎么回事？说来听听。

小姚：我上周跟一分公司的员工聊天，谈到收入的时候，发现他们对目前的收入不太满意。于是我与他们的主管谈了一下，发现的确存在一些问题。我总结了一下，主要有这些问题。

(1) 认为公司在薪酬调整方面比较随意。每年工资调整的时候，调整幅度没有明确标准，主要取决于部门经理或者公司领导的主观感受。

(2) 工资没有很好地体现业绩差异。部分员工的工资基本上是固定发放，与员工的工作表现、实际努力脱节，干多干少都一样。另外有几个员工反映绩效考核不合理，业绩较好的员工与资格老但业绩普通的员工收入差不多，甚至还低一些，感到不公平。

(3) 没有很好地体现岗位特点。有业务主管抱怨，同样是主管，固定工资却不一样，入职时议价能力越强，入职后的固定工资就越高，而不是基于工作岗位和性质来决定。

(4) 奖金发放缺乏透明性。员工的年终奖占奖金的绝大部分，但对年底能拿多少员工心里没底，并且年终奖是保密的。

(5) 部门之间薪酬不平衡。一些部门的员工反映，业务部门和支撑部门之间的薪酬差距过大，容易导致非业务部门的员工有不满情绪。

(6) 员工工资与同行业其他公司的员工相比，缺乏竞争性，工资水平偏低。

Miss 陈：总结得很好，很细致。你说的情况，涉及薪酬管理的许多方

面，其中之一是公平性问题，包括了内部公平性、个人公平性和外部公平性。咱们就薪酬公平性的问题做深入的探讨吧，我把你谈的情况归类整理一下，如图 5-1 所示。

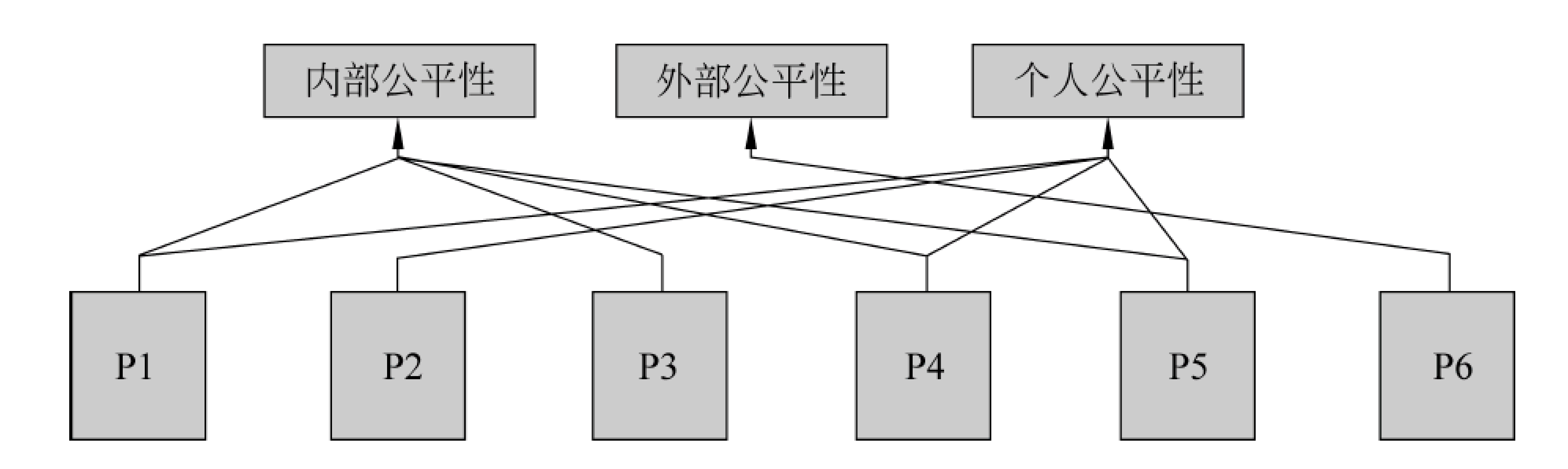


图 5-1 薪酬公平性的分解示意图

小姚：是的，经理，这些问题影响了员工的薪酬公平感，导致薪酬满意度较低，工作积极性受到一些影响。

Miss 陈：那么你觉得我们该做些什么呢？

小姚：我认为应该对公司的薪酬现状进行盘点分析，然后再研究制定优化、改进措施。毕竟，目前这只是某个分公司的个别现象，不能代表整个公司的情况，其他分公司或部门有没有类似情况，还需要进行调查才能确定。而且这是通过谈话得到的信息，主观性较强，没有数据做支撑，难以判断真实情况。

Miss 陈：其实从人性角度来看，人对财富的欲望和需求是一直存在的，所以通常企业员工对薪酬的满意度不会很高，会认为收入应该再多些。即使是一些薪酬水平很高的企业，它们的员工薪酬已经高于社会水平、行业水平很多了，但仍然难以完全满足员工对薪酬的欲望。因此员工口头反映的薪酬问题，需要深入调研和分析，以实际情况为依据，避免出现偏差。

小姚：是啊，不能道听途说，咱们得仔细分析，看看是否真的有问题。不过，该怎么分析呢，要不要做一次薪酬满意度问卷调查，通过问卷来收

集数据进行分析呢?

Miss 陈:可以进行问卷调查,不过在这之前,我们可以利用现有的薪酬数据,开展一些关于薪酬公平程度方面的数据分析,对薪酬的公平性进行总体的了解和把握,然后再进行问卷调查,效果会好不少。

小姚:通过对现有数据的分析就能知道薪酬公平性的情况吗?

Miss 陈:是的,有一些技术可以从总体上对薪酬公平程度进行分析,比如用薪资结构图法、基尼系数、薪酬公平感计量模型等,下面我们来分别讲讲。

小姚:好的。

5.2 分析方法

5.2.1 薪资结构图

小姚:薪资结构图是不是将公司的薪酬结构用图表的方式展示出来?

Miss 陈:是的。我们公司实行的是组合薪资结构,包括岗位工资、能力工资和绩效工资三个部分。其中能力工资以能力为导向,与工龄、学历、职称、职业资格、持证等因素挂钩,反映员工的能力水平,这部分薪酬相对比较固定;绩效工资与员工的实际工作表现挂钩,反映员工的业绩水平,这部分薪酬是浮动的;岗位工资与担任的职务的重要程度、任职要求和劳动环境对员工的影响挂钩,主要受岗位等级影响,并且对能力工资和绩效工资都有影响,岗位等级越高的员工,能力工资和绩效工资相应会比较高。因此,按照我们公司设置的 12 级岗位等级,合理的薪资结构图如图 5-2 所示。

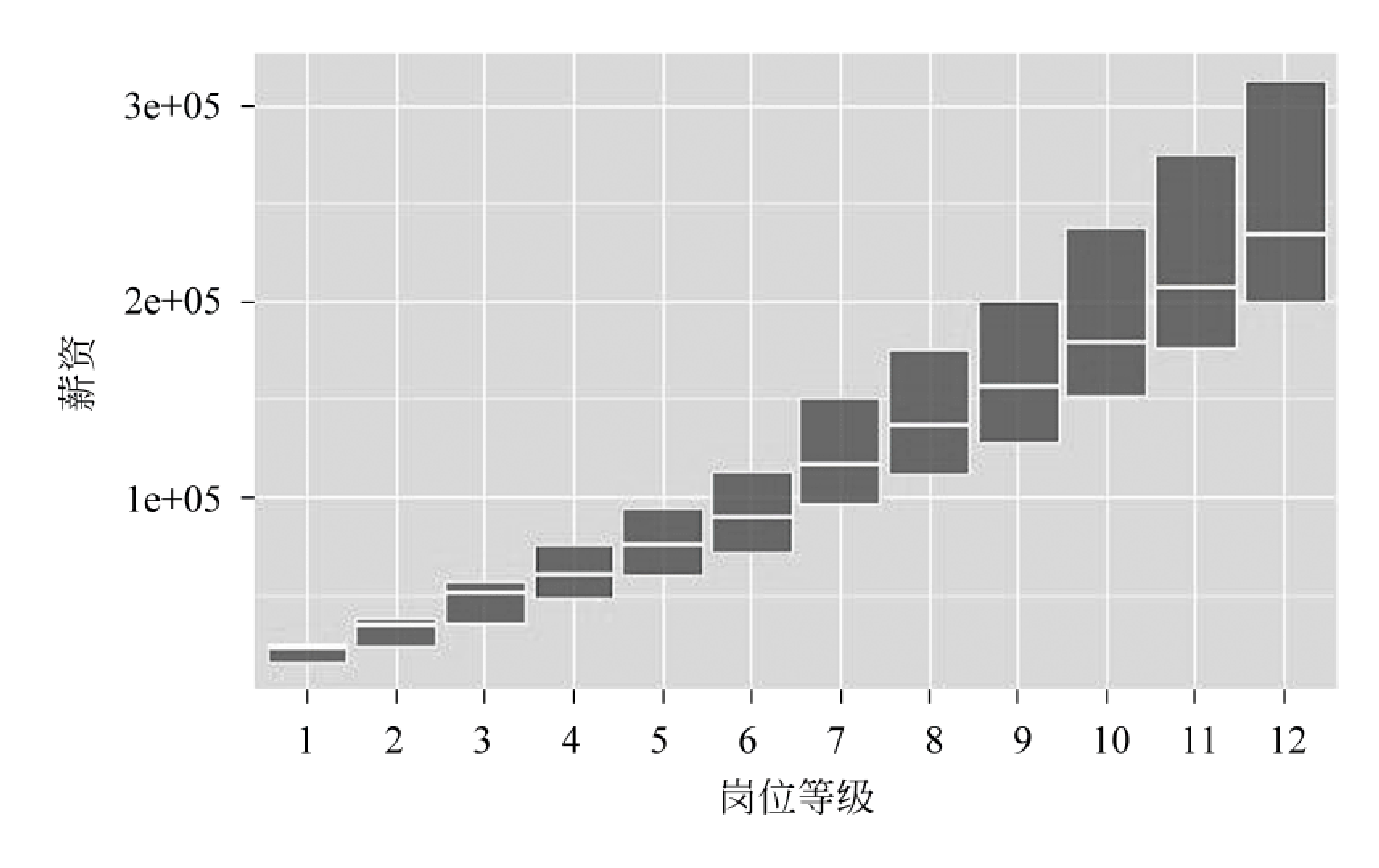


图 5-2 合理的薪资结构图

薪资结构图的 R 语句如下：

```
library(ggplot2)
d<-read.csv("第五章/薪资结构.csv")
d$岗位等级<-factor(d$岗位等级,levels=rev(d$岗位等级),ordered=T)
g<-ggplot(d)
g+geom_crossbar(aes(岗位等级,薪资,ymin=最小值,ymax=最大值),
fill="blue",alpha=0.7,colour="white")+
  labs(title="薪资结构图")
```

小姚：这个薪资结构图我学过呢，图中反映了很多和薪资有关的信息。

Miss 陈：你说说看，反映了哪些信息？

小姚：比如薪酬级差，就是不同等级之间薪酬相差的幅度。从图上可以看出，每个岗位等级之间的薪酬是有差距的，这种差距不能太大，否则会造成员工不团结，也不能太小变成吃“大锅饭”而使员工没有积极性。并且相邻层级之间应该有一定程度的重合，即低一级岗位的员工若做得好，可以获得高一级岗位中等程度左右的薪酬。当然也需要充分考虑等级之间在劳动强度、复杂程度、责任大小方面的差别，以达到激励的目的。

由于岗位级别越高，岗位之间的劳动差别越大，工作价值差别越大，所以，高级别岗位之间的薪酬级差要大一些，低级别岗位之间的薪酬级差要小一些。

Miss 陈：说得很好。图 5-2 是标准的薪资结构图，代表了合理的薪资等级分布结构，基于标准的薪资结构图，我们通过观察公司实际的薪资结构图来进行对比，就可以分析薪资结构是否合理了。

小姚：明白了，就是通过对比分析的方法，用实际薪资结构图与标准薪资结构图进行对比来分析薪酬设置的合理性。

5.2.2 基尼系数

Miss 陈：还可以用基尼系数来分析薪资公平性。关于基尼系数，你知道什么吗？

小姚：基尼系数大名鼎鼎，是用来反映收入差距程度的。我经常在网站、报纸上看到相关报道会披露一些国家的基尼系数，用来反映国家内部的贫富差距。

Miss 陈：那么你说说什么是基尼系数？

小姚：基尼系数是 1943 年美国经济学家阿尔伯特·赫希曼根据洛伦茨曲线所定义的，判断收入分配公平程度的指标。它是一个比例数值，在 0 和 1 之间，是国际上用来综合考察一个国家居民内部收入分配差异状况的重要分析指标。

说到基尼系数就必须提到洛伦茨曲线。如图 5-3 所示，在洛伦茨曲线中，若设实际收入分配曲线和收入分配绝对平等曲线之间的面积为 A（图中灰色阴影部分），实际收入分配曲线右下方的面积为 B，那么可以用 A 除以（$A+B$）来表示收入的不平等程度，这个数值被称为基尼系数或称洛伦茨系数。公式很简单：

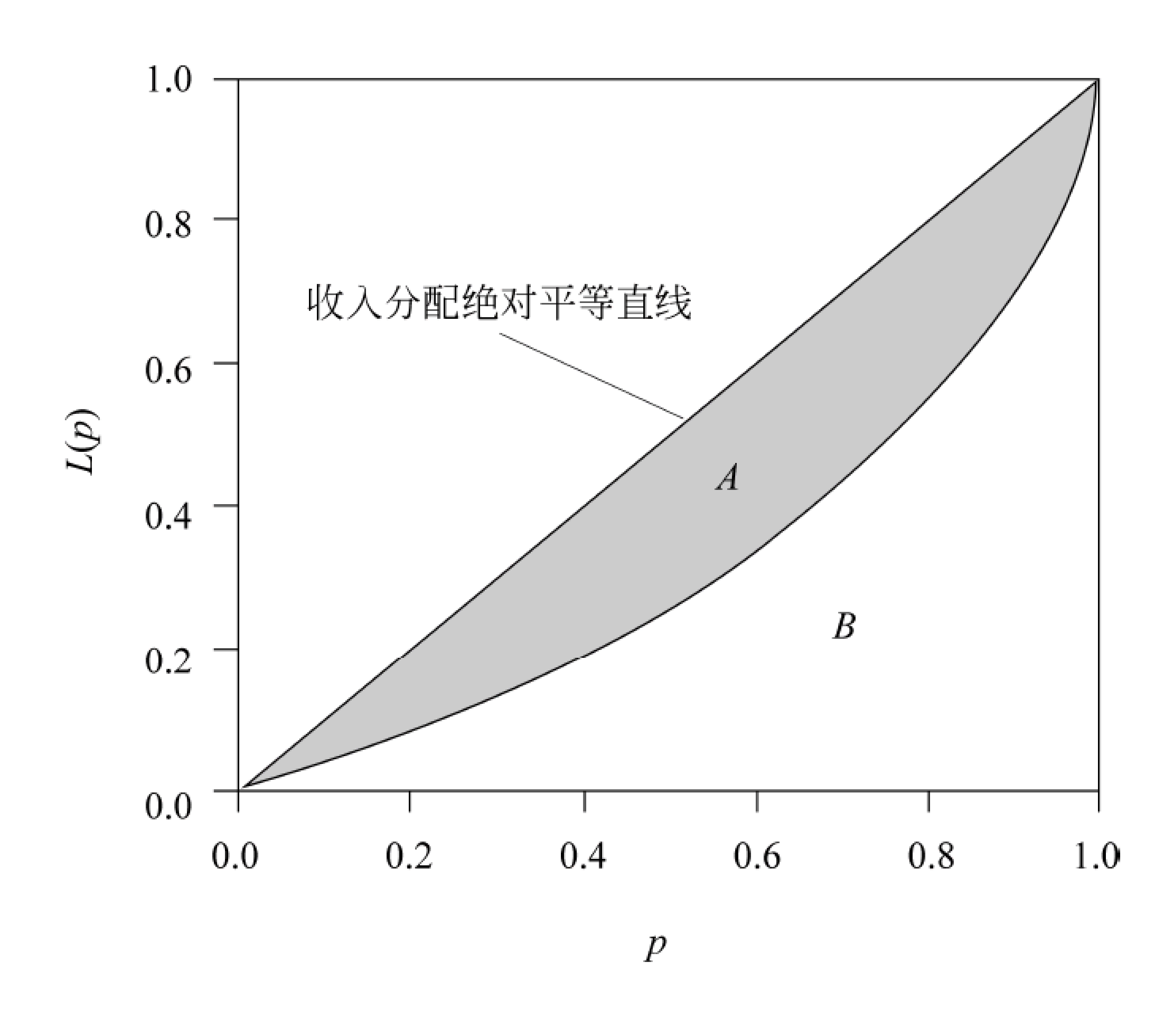


图 5-3　洛伦茨曲线

$$G = \frac{A}{A + B}$$

如果 A 为零，基尼系数为零，表示收入分配完全平等；如果 B 为零则系数为 1，表示收入分配绝对不平等。收入分配越趋向平等，洛伦茨曲线的弧度越小，基尼系数也越小；反之收入分配越趋向不平等，洛伦茨曲线的弧度越大，那么基尼系数也越大。

Miss 陈：很好，用基尼系数来判断收入差距的标准是什么呢？

小姚：通常把 0.4 作为收入分配差距的“警戒线”。基尼系数大于等于 0.4 说明收入差距较大，容易引起不公平的感觉，容易出现社会震荡；而基尼系数小于 0.4 则说明收入差距相对合理或平均，具体见表 5-1。

Miss 陈：你平常看到的基尼系数多是用来表示一个国家的贫富差距程度吧，有没有看到过在企业内用基尼系数的情况呢？

小姚：没有。根据我查到的资料，基尼系数通常是用于反映一个国家的收入分配情况，没怎么听说过企业用基尼系数的情况。

表 5-1　基尼系数各分值及其意义

基尼系数	意　　义	基尼系数	意　　义
低于 0.2	收入绝对平均	0.4～0.5	收入差距较大
0.2～0.3	收入比较平均	0.5 以上	收入差距悬殊
0.3～0.4	收入相对合理		

Miss 陈：其实企业也可以用基尼系数来反映企业内部的收入差距①，甚至可以用来表达任一群体、任一事情上的差异程度，不限于经济，不限于收入。企业基尼系数有不少优点。

(1) 用一个数值就可以反映总体的薪酬差别，这对研究经营管理人员和职工收入增长的关系是十分必要的。

(2) 基尼系数是国际经济学界常用和成熟的度量指标，也是我国常用的经济指标，具有较高的信度和效度，相对容易理解。

(3) 基尼系数的计算比较简便。计算基尼系数的方法有十几种，常用的四种计算方法有：直接计算法、拟合曲线法、分组计算法和分解法。引用百度百科上的一个计算公式②如下：

$$G = 1 - \frac{1}{n}\left\{2\sum_{i=1}^{n-1} w_i + 1\right\}$$

小姚：经理，是不是对企业来说，基尼系数的大小就反映了企业内部薪酬差距的大小？

Miss 陈：是的。如果企业内部的基尼系数过大，那么员工感觉薪酬差距过大，就会产生较强烈的不公平感。系数越大，这种不公平感觉就越

① 王今舜，马彤. 运用基尼系数增强企业薪酬制度的公平性[J]. 经济与管理研究，2008(2)。

② 引用自百度百科，公式含义：假定一定数量的人口按收入由低到高排队，分为人数相等的 n 组，从第 1 组到第 i 组人口累计收入占全部人口总收入的比重为 w_i，则说明：该公式是利用定积分的定义将洛伦茨曲线的积分(面积 B)分成 n 个等高梯形的面积之和得到的。

强烈，导致人心不稳，最终影响企业的经营管理。

小姚：明白了，但是怎么计算企业的基尼系数呢？

Miss 陈：关于计算的问题，我们等一下结合实际数据再说吧。

5.2.3 薪资均衡指标 Compa

Miss 陈：还有一个专门用于衡量薪酬均衡程度的统计量，叫薪资均衡指标，也叫 Compa 系数。

小姚：什么是薪资均衡指标呢？

Miss 陈：薪资均衡指标是一个衡量和评估酬薪体系的统计量，是一个相对指标，既可以检测员工个人的薪酬水平是否均衡，也可以检测部门的薪资均衡程度，还可以检测公司在行业中的薪资均衡程度。所以，薪资均衡指标广泛应用在人力资源管理的薪资制度诊断和管理中，用于检测薪酬分布是否均衡，是人力资源管理中一个有力的计划和控制工具。

小姚：这个指标计算起来应该很复杂吧？

Miss 陈：不然，薪资均衡指标计算起来比较简单，它是平均数和中位数的比值，是一个相对数，计算公式如下：

$$\text{Compa}=\frac{\text{薪资平均值}}{\text{薪资中位数}}$$

(1) 当用于计算个人薪资均衡指标时，反映的是单个员工的工资相对部门或者企业工资范围中位数的比例，这种情况下，公式中的分子就是个人的工资数，公式如下：

$$\text{Compa}_{\text{个人}}=\frac{\text{个人实际所得薪资}}{\text{部门或企业薪资中位数}}$$

(2) 当用于计算部门薪资均衡指标时，反映的是该部门人员工资与企业工资范围中位数的比例，这种情况下，公式中的分子就是该部门员工的平均工资，公式如下：

$$\text{Compa}_{\text{部门}}=\frac{\text{部门平均薪资}}{\text{企业薪资中位数}}$$

(3) 当用于计算企业在行业中的薪资均衡指标时,反映的是企业的工资水平在行业中的情况,这种情况下,公式中的分子就是企业的人均工资,分母则是人才市场中行业酬薪的中位数,公式如下:

$$\text{Compa}_{\text{企业}}=\frac{\text{企业平均薪资}}{\text{行业薪资中位数}}$$

姚:薪资均衡指标的计算不算复杂,不过结算结果怎么应用呢?

Miss 陈:主要应用于分析薪资的均衡程度,了解薪酬水平在群体中处于什么位置。比如对员工个人来说,当其薪资均衡指标大于等于 1.0 时,表明总体上员工被支付了等于或高于他们工资范围中位数的工资。对胜任岗位的员工来说,应该支付等于或高于中位数的薪资。

而当薪资均衡指标低于 1.0 时,则说明员工工资偏低,低于他们工资范围中位数的工资。出现这种情况要分析原因,可能的原因有:员工个人能力不胜任工资岗位、工作绩效偏低、工龄短、学历低等。针对不同的原因要找到解决问题的方法,比如安排培训、加强激励、鼓励参加学历教育,等等。

在运用薪资均衡指标进行分析时,最好分析同类人员,比如分析岗位等级相同、部门相同、专业相同的员工,这样才会具有较高的可比性。岗位等级、部门、专业等因素对薪酬都有较强的影响,比如高岗位的人员薪酬一定比低岗位的高,若放在一起比较,必然出现低岗位人员的薪酬均衡指标偏低的现象。

小姚:这个指标感觉很实用,一定得试试。

5.2.4 公平感计量模型

小姚:经理,前面的方法都是从总体上去把握薪酬的公平、均衡程

度，但如果要准确衡量员工个人的薪酬公平感，该怎么办呢？是不是要做薪酬满意度调查呢？

Miss 陈：不一定需要薪酬满意度调查。这种调查涉及范围比较大，薪酬公平性只是其中的一个维度而已。通常薪酬满意度调查除了调查薪酬公平性，还会调查薪酬制度执行情况、薪酬福利激励效果、公司可信度等内容，以及一些心理层面的主观感受，还会向员工征求意见，调查的内容和范围都比较大。并且最重要的是薪酬满意度调查也是对总体的分析，很少分析个人。

小姚：那该怎样才能比较准确地了解某个员工的薪酬公平感呢？

Miss 陈：可以尝试使用薪酬公平感计量模型①来分析员工个人的薪酬公平感。

小姚：没听说过呢，这个薪酬公平感计量模型是什么呢？

Miss 陈：在介绍这个分析模型之前，我们先讨论一下薪酬公平感的相关内容。通常我们认为一件事情是否公平，会考虑两个要素，你知道是哪两个要素吗？

小姚：我想应该是投入和回报吧。

Miss 陈：很好，投入和回报的情况直接影响我们对公平感的认知。如果投入大回报小，就会觉得吃亏，心理不平衡，公平感降低；如果投入和回报相当，就会觉得公平；如果投入小回报大，就会觉得非常满足，公平感上升。

小姚：感觉和企业的经营管理有相通之处，如果企业投入的成本和收益持平，就是盈亏平衡；如果成本大过收益，就是亏损；如果收益大于成本，就是盈利。

Miss 陈：是的，你说对个人而言，如果要获得薪酬上的公平感，是否也要考虑投入和回报呢？

---

① 周霞，李国辉，石爱玲. 薪酬管理中的公平感计量模型[J]. 中国管理科学，2005(10).

小姚：是的。

Miss 陈：那么请再想想看，对个人而言，投入是什么，回报又是什么呢？

小姚：我想，投入就是工作上付出的劳动，回报就是薪酬，是吗？

Miss 陈：你说的是直接的投入和回报，范围小了点。按薪酬公平感计量模型的研究，员工的投入包括了“员工身上的知识、技能和健康在工作中的资本化，其投入量还要受到自身努力程度和工作任务的影响”，所以投入应该包括五个要素：技能、知识、健康、任务和努力。其中一些要素在员工工作之前就已经投入了一部分，比如技能、知识等，这些投入可用学历、证书等来反映，都是员工对自身的投入，需要且应该资本化。

回报则不仅仅是薪酬。从薪酬的全面性来看，包括四个要素：现金薪酬、福利、培训和晋升机会。另外在工作过程中，员工的知识和技能在提升，这些在工作过程中提升的知识和技能也应计入工作回报。

由于公平感是员工经过横向和纵向比较之后产生的一种主观的、相对的感觉，所以如果能计算出投入和回报的比率，就能比较恰当地用量化的数据来反映公平感。这个比值就是公平感比率，计算公式如下：

$$E_{\text{公平感比率}}=\frac{Q_{\text{薪酬}+\text{福利}+\text{培训}+\text{晋升}+\text{知识}+\text{技能}}}{I_{\text{知识}+\text{技能}+\text{健康}+\text{任务}+\text{努力}}}$$

小姚：那么是否可以根据上面的公式得出如下推论：如果 $E\approx0$ 就表示员工对薪酬基本满意，在薪酬上具有公平感知；如果 $E<0$ 就表示员工对薪酬不满意，薪酬公平感低，E 越小公平感越低；$E>0$ 就表示员工对薪酬满意，E 越大满意度越高。

Miss 陈：是的。

小姚：那么这九个要素之间的权重怎么确定呢？

Miss 陈：权重分配可用层次分析法来计算。

小姚：看来要学不少东西呢，能介绍一下层次分析法吗？

Miss 陈：层次分析法是一种决策方法。比如你要买一辆车，经过初步筛选，看上了三个车型，各有优缺点，犹豫不决到底该买哪一辆，这时就可以使用层次分析法来帮助你做决定。层次分析法可以将主观判断通过两两比较的方式转换为量化数据，从而计算每个车型的权重。根据计算结果，权重最高的那个车型就是你最优的选择。

小姚：这个方法很实用呢，可以用来选车，等我买车的时候就试一下。

Miss 陈：你学习了之后，不仅买车时可以用，在很多需要决策的地方都可以使用。层次分析法可以将主观判断量化，从而选择最优方案。这个方法是美国运筹学家匹茨堡大学教授萨蒂于 20 世纪 70 年代初，为美国国防部研究“根据各个工业部门对国家福利的贡献大小而进行电力分配”课题时，应用网络系统理论和多目标综合评价方法，提出的一种层次权重决策分析方法。

小姚：比较好奇为什么叫层次分析法呢？

Miss 陈：因为该方法通常会把要分析的元素分成目标层、准则层、方案层等层次，然后在此基础上采用定性和定量结合的方法来进行分析和决策。接着刚才的例子，比如你要买一辆车，那么“买车”就是目标，这属于目标层；经过你的初步筛选，选出了三个车型，而你要在这三个车型中选择一个购买，那么这三个车型就叫作备选方案，属于方案层；而你在选购一辆车的时候，通常会考虑一些要素，比如外观、性能、价格、大小、排量、用途等，会根据这些要素来综合判断，那么这些要素就是判断的准则，属于准则层。

小姚：哦，原来层次是指目标层、方案层和准则层。但是层次分析法是如何将主观的判断转换为量化数据的呢？

Miss 陈：层次分析法先会根据重要性对备选方案进行两两比较，计算重要性差异，用数据来表示。比如方案 1 比方案 2 重要，就记为 1，同等重要，就记为 0，不太重要，就记为 −1，通过这种方式将主观判断进行量

化。各个方案完成两两比较后，就会构成一个判断矩阵，然后计算这个矩阵的最大特征根和对应的特征向量，再把特征向量归一化后即为权重。

小姚：两两比较构建判断矩阵这个能理解，但是最大特征根和对应的特征向量，又要怎么计算的呢？

Miss 陈：具体计算可借助专业的层次分析法软件①，我们不用手动计算。通常我们只需要将精力放在方案和准则之间的两两比较上，重点关注主观判断量化的过程，其余的计算过程交给软件去执行即可。软件会自动计算最大特征根和特征向量，最终计算出各个方案的权重。

好了，回到我们的主题薪酬公平感计量模型。研究者使用了层次分析法来进行分析，构建了以下三个层次，如图 5-4 所示。

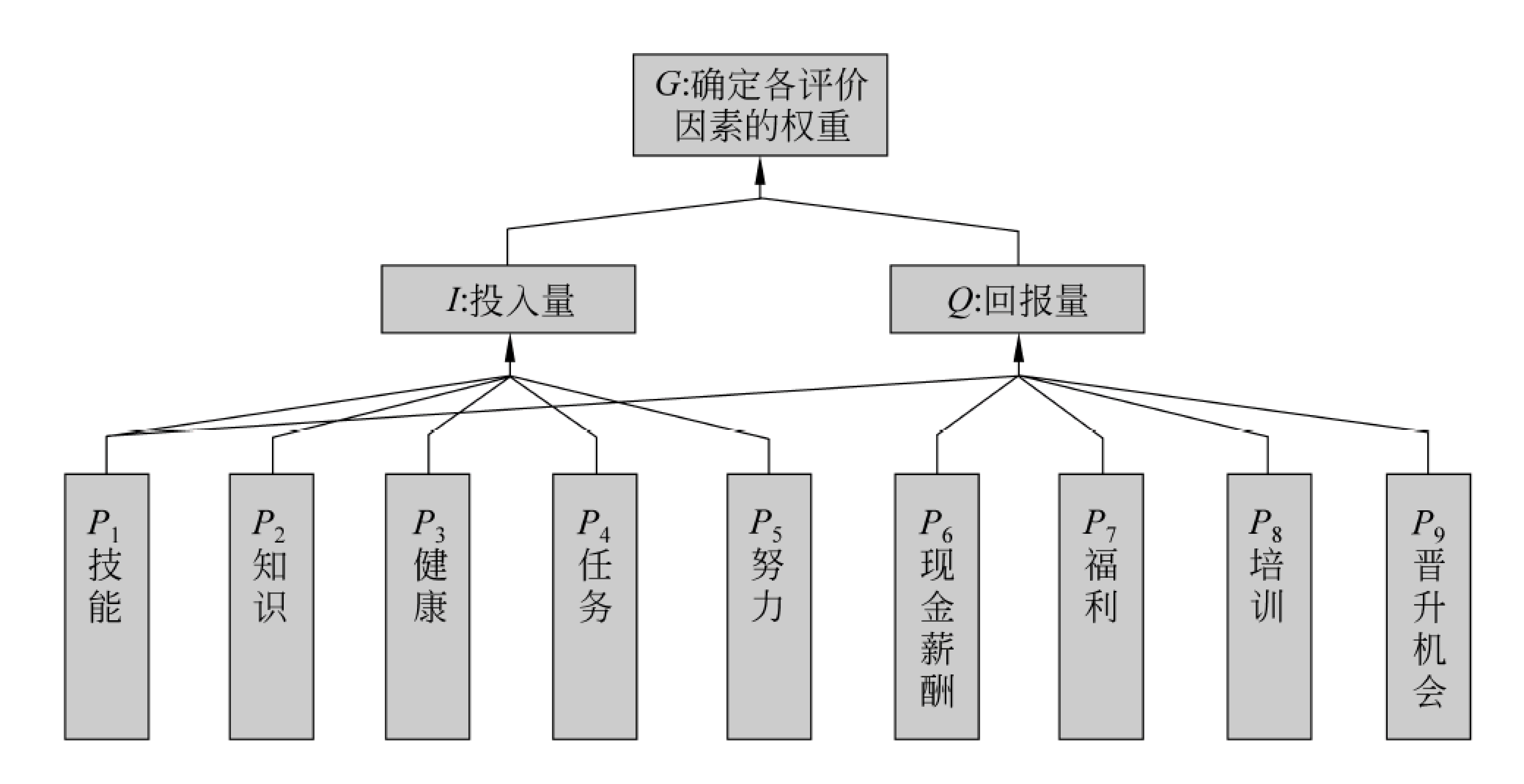


图 5-4 薪酬评价因素的层次分析

（1）目标层：确定各评价因素的权重。

（2）准则层：投入量、回报量。

（3）方案层：技能、知识、健康、任务、努力、现金薪酬、福利、培训、晋升机会。

① 推荐使用层次分析法专用软件 yaahp 10 以上版本，该软件基本功能免费。

研究者同时也计算出了各个要素的权重，具体见表 5-2。

表 5-2　薪酬评价要素的权重

类型＼要素		技能	知识	健康	任务	努力	现金薪酬	福利	培训	晋升机会
投入量	0.45	0.47	0.09	0.04	0.14	0.26				
回报量	0.55	0.06	0.04				0.42	0.24	0.09	0.15

小姚：看上去，投入量和回报量的权重不相同？

Miss 陈：是的，回报量的权重略大些。从权重表上可以看出，对投入影响最大的要素依次是：技能、努力、任务，合计权重超过 80%；对回报影响最大的要素依次是：现金薪酬、福利、晋升机会，合计权重超过 80%。这和我们的日常理解吻合。

小姚：要素权重是怎么计算出来的呢？

Miss 陈：刚才说了，这是通过层次分析法建立的判断矩阵计算出来的。至于计算的具体过程和方法，如果你想手动计算，可以看看有关矩阵运算方面的内容，学习如何计算最大特征根；如果你想把重点放在确定准则层和方案层之间的关系、如何构建判断矩阵等方面的话，直接使用软件即可。其实各个要素的权重已经计算出来了，我们可以直接使用，不一定要重新去计算。

小姚：哈哈，还是用软件计算效率高啊。

5.3 数据准备

Miss 陈：你准备一些薪酬数据吧，我们结合实际数据来看看如何应用这些方法。

小姚：好的，我整理了公司去年每个员工的薪酬数据[①]，具体见表 5-3。

表 5-3 员工薪酬数据示例

部 门	部门类型	姓 名	岗位等级	全年收入(元)
经理室	管理者	员工 927	11	96 000
经理室	管理者	员工 867	11	95 982.5
企业发展部	职能部门	员工 33	9	95 876.5
项目管理二部	技术部门	员工 51	9	95 584
经理室	管理者	员工 863	11	94 823
企业发展部	职能部门	员工 32	9	93 860
管线部	生产部门	员工 1 020	2	93 159
经理室	管理者	员工 855	11	92 499
综合部	职能部门	员工 38	9	91 572.4
经理室	管理者	员工 182	11	91 341
事业部	生产部门	员工 659	6	91 200
经理室	管理者	员工 135	10	90 487
管线部	生产部门	员工 634	7	89 664
管线部	生产部门	员工 1 006	4	85 426.3
……	……	……	…	…

Miss 陈：有了这些数据，我们就可以进行计算和分析了。

① 本数据纯属虚构。

5.4 分析过程

5.4.1 用薪资结构图分析薪酬结构合理性

Miss 陈：首先我们用薪资结构图法来进行分析。

小姚：经理，我根据上一年度实际的薪资数据绘制了薪资结构图，请您看看。由于同一岗位等级的薪资极差比较大，所以我用了箱型图来表示，如图 5-5 所示。

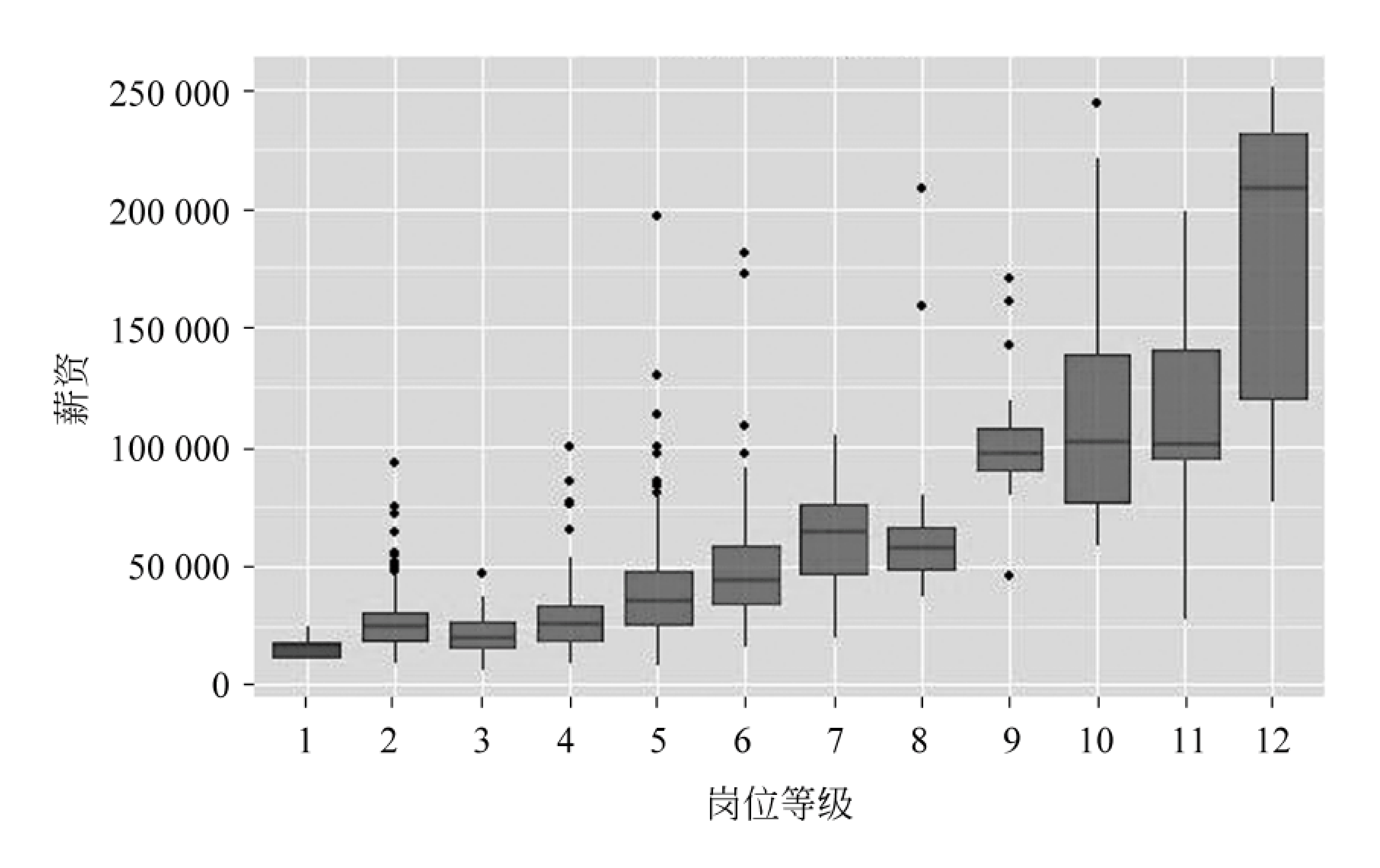


图 5-5　公司实际薪资结构图(各岗位层级箱形图)

薪资结构图实(实际)的 R 语句如下：

```
library(ggplot2)
d<-read.csv("第五章/薪酬分析.csv")
d$岗位等级 <-factor(d$岗位等级,levels=rev(d$岗位等级),ordered=T)
g<-ggplot(d)
g+geom_boxplot(aes(岗位等级,应发工资),fill="blue",alpha=0.7)+
```

```
labs(title="薪资结构图(实际)",y="薪资")
```

Miss 陈：有了实际的薪资结构图，对比标准薪资结构图，我们就可以进行比较分析了。分析薪资结构图要关注和解答以下问题。

(1) 各岗位等级之间是否保持一定的级差？是否岗位等级越高薪资越高？

(2) 每个岗位等级的薪酬是否保持一定级差？随着岗位等级上升级差范围是否逐渐加大？

(3) 是否存在低岗位高薪酬、高岗位低薪酬的现象？

小姚：我试着分析一下。

(1) 从实际薪资结构图可以看出，各岗位等级之间保持了一定的级差，且岗位等级越高，薪资就越高，级差也越大。不过也出现了一些例外情况，比如二岗薪酬高于三岗、八岗薪酬低于七岗、十一岗薪酬略低于十岗，有些异常。

(2) 每个岗位等级都有一定的薪酬幅度，而且随着岗位等级的上升，薪酬幅度的变化范围在增加。但是八、九、十一岗的薪酬幅度比较小。

(3) 存在低岗位高薪酬、高岗位低薪酬的现象。比如二、四、五、六、八、九、十岗，有少数人的薪酬远远高出同岗位其他人员，甚至超过高几个层级人员的最高收入，而九岗则出现一个员工收入过低的现象(图中的小黑点表示异常值)。

Miss 陈：很好，不过这是对总体情况的分析，发现状况时还需要进一步细化分析。你知道薪酬不仅受到岗位等级的影响，还受到岗位性质、技能水平等因素的影响，比如你发现存在低岗位高薪酬现象，那么还需要进一步研究岗位性质。对于市场营销、项目管理类岗位，由于这两类岗位人员的收入和业绩紧密挂钩，绩效工资占薪酬比例大，极有可能出现低岗位高薪酬的现象，这种情况的低岗位高薪酬现象是合理的，也是可以解释

的。但如果是一些职能、技术类岗位出现了低岗位高薪酬的现象，就需要引起注意了。

小姚：嗯，明白了。用薪资结构图法的方式可以很直观地看到各岗位等级的薪酬分布，以及各个岗位等级的薪酬水平、级差、分布、异常值等情况，这对研究和分析薪酬的总体情况、薪酬机制的合理性非常有用。看到问题后，结合进一步的研究分析，就能找出原因并制定调整措施了。

5.4.2　用基尼系数分析总体薪酬差距

Miss 陈：接下来我们看看如何计算基尼系数。实际上，用 R 语言可以很方便地计算基尼系数，并且绘制洛伦茨曲线。

小姚：好的，我来试试吧。根据去年的薪酬数据，我用 R 语言的 ineq 包中的 Gini 函数计算基尼系数，计算结果如下：

$$G_{公司}=0.38$$

同时绘制洛伦茨曲线，如图 5-6 所示。

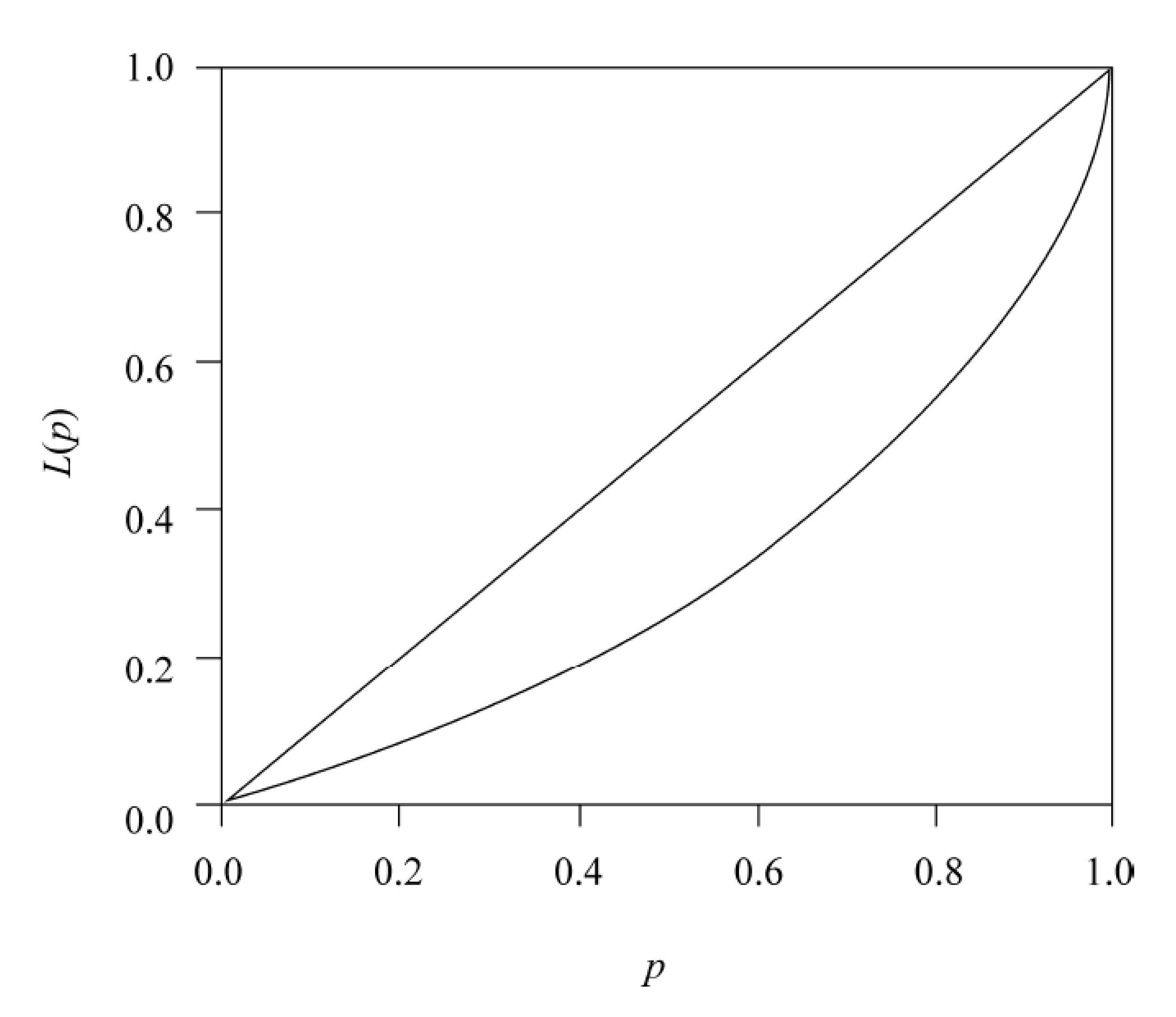


图 5-6　谦多顺公司的薪酬洛伦茨曲线

所用 R 语句如下：

```
library(ineq)
d<-read.csv("第五章/薪酬分析.csv")      #读取数据
Gini(d$应发工资)                         #计算基尼系数
plot(Lc(d$应发工资),main="谦多顺公司薪资洛伦茨曲线",col= 2)
                                         #绘制洛伦茨曲线
```

Miss 陈：很好。从计算结果看到基尼系数等于 0.38，说明公司的总体薪酬差距并不算大，收入差距相对合理。虽然你发现了部分员工反映收入差距大，觉得不公平，但从总体上看并不是这样，说明这可能只是个别分公司中出现的个别现象。倒是部分员工提到收入没有按业绩体现差异，干多干少一个样，这种现象需要关注。

小姚：看来经过数据分析之后，我们对公司的薪酬现状了解得更清楚了。不过经理，如果基尼系数计算出来的结果大于 0.4，那么该怎么办呢？

Miss 陈：如果基尼系数计算出来的结果大于 0.4，是不是就不好呢？是不是需要进行调整呢？其实这方面还没有定论，要结合企业实际情况来判断。如果公司处于高速发展期、创业期等发展阶段，需要快速开拓市场、创造利润，就需要加大员工激励力度，将有限的人工成本投入核心人员身上，给核心人员高激励性薪酬，拉开核心人员和普通人员的薪酬差距，这种情况下基尼系数就会偏大，但这对企业发展是有利的。如果企业处于稳定发展期，以员工队伍稳定为主要目标，就需要体现薪酬公平性，提高员工薪酬公平感，这种情况下基尼系数就不宜偏大。

此外，企业的类型也会影响基尼系数。比如研发型企业，高级知识分子多，产品开发周期长，为保证研发队伍稳定，员工之间的薪酬差距不宜过大，这种情况下基尼系数会偏低。而对于一些互联网企业，为追求快速发展，对员工的激励程度很大，特别是能带来业绩的员工，其薪酬可能会比其他员工高数倍或数十倍，这种情况下基尼系数则偏高。

国有企业受到工资总额的限制，基尼系数倾向于偏低，而民营企业和外资企业不受此限制，基尼系数倾向于偏高。实际上一些知名企业的基

尼系数会偏高，特别是一些国外上市企业，其中高层管理人员的薪酬与普通员工的薪酬差距非常大，加上股票、期权等高管专享的长期性激励措施，这类企业的基尼系数可能会相当高。但不能凭基尼系数说这类企业的薪酬不合理、员工薪酬满意度低、公平感低、企业不稳定，相反这些企业可能发展得相当好。

小姚：明白了，就是说要结合企业的实际情况来理解基尼系数。

Miss 陈：是的，如果发现基尼系数高，那么首先要分析企业的实际情况。确属异常时，为促进企业内部分配差距在合理范围之内，可以考虑进行薪酬调整，将基尼系数定位到合理范围。确定基尼系数调整目标后，可以倒推高层管理人员的薪酬标准，重新调整薪资结构。

5.4.3　用薪资均衡指标分析各岗位薪资均衡程度

小姚：经理，我来计算一下薪资均衡指标 Compa 系数吧。先试试计算个人 Compa 系数。找个部门试试，哈哈，就拿财务部试验吧。计算结果如图 5-7 所示。

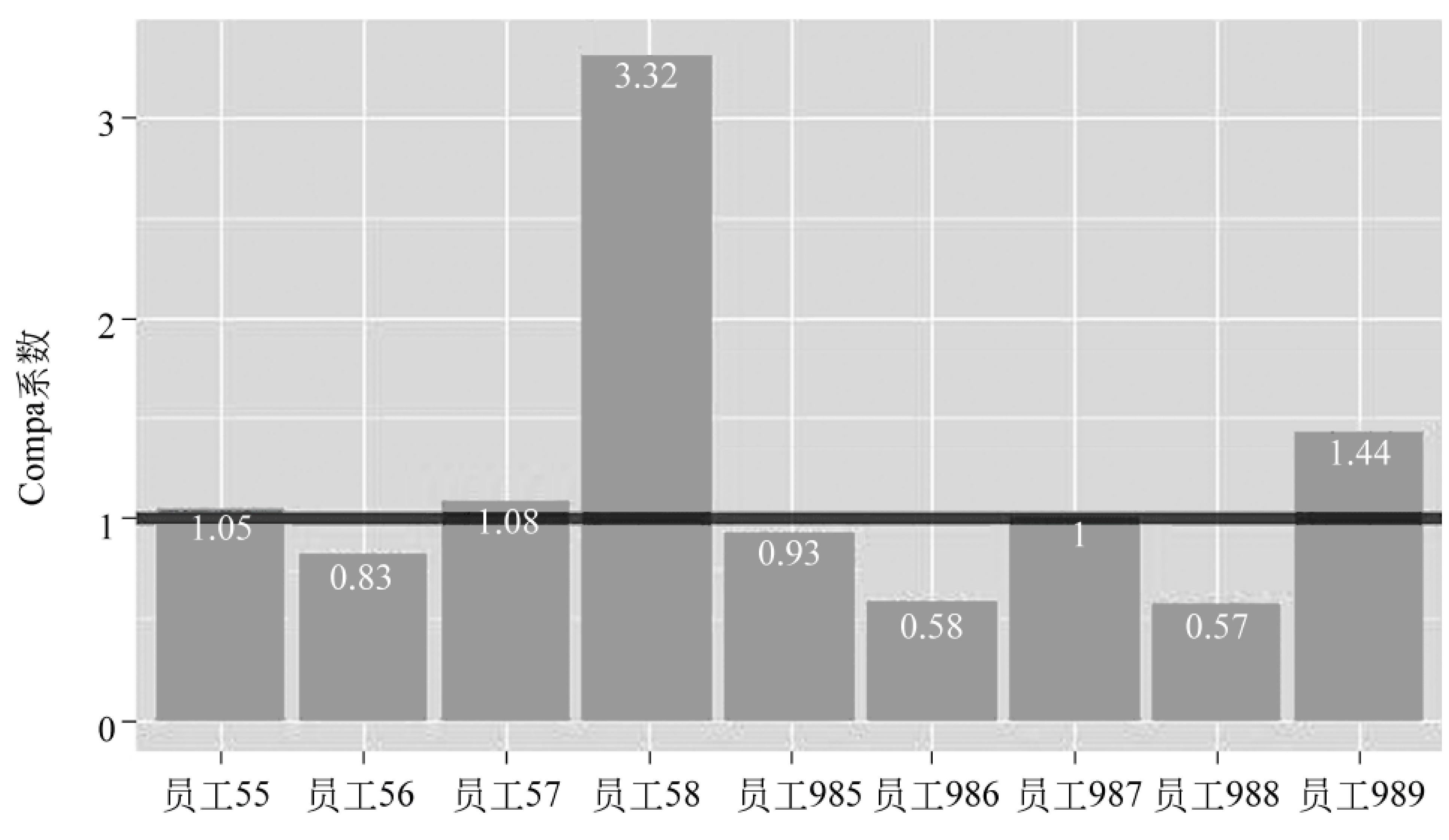


图 5-7　谦多顺公司财务部员工 Compa 系数

R 语句如下：

```
library(ggplot2)
d<-read.csv("第五章/薪酬分析.csv")
d<-d[d$部门=="财务部",]
d$compa<-d$应发工资/median(d$应发工资)
g<-ggplot(d)
g+geom_bar(aes(姓名,compa),stat="identity",fill="red",
alpha=0.7)+geom_hline(size=2,yintercept=1,colour="blue",
alpha=0.7)+geom_text(aes(姓名,compa,label=round(compa,2)),
vjust=1,colour="white")+labs(title="财务部员工 compa 系数",x="")
```

根据计算结果可以看到，58 号员工的 Compa 系数最高，55、57、989 号员工的 Compa 系数大于 1，说明其薪酬水平偏高，987 号员工的 Compa 系数等于 1，正好均衡，其他人员的 Compa 系数都小于 1。计算结果和实际情况比较一致，58 号员工是财务部经理，55、57、989 号员工是财务部的三位骨干人员，987 号是一个老员工，其他几人是新来的人员。

Miss 陈：很好。

小姚：如果用薪资均衡指标来分析部门的薪酬会是怎样的情况呢？

Miss 陈：如果把薪资均衡指标用来衡量部门或者分公司的薪酬水平，那么主要是用来考察不同部门和分公司之间的薪酬水平是否公平，企业的薪资差距是否与企业的战略相匹配。比如，通过薪资均衡指标可以考察不同部门之间的薪酬水平，结合分析部门之间的薪酬差距，从而判断部门之间的薪资差距是否合理，是否符合部门的价值权重。分析同类部门之间的薪资差距，还可以研究出现差距的原因。是人为误差（有些部门之间考核一团和气，人人得高分），还是部门业绩差异所致。

小姚：明白了。现在我再来计算一下各个部门的薪资均衡指标吧，选一些职能部门来试试看。计算结果如图 5-8 所示。

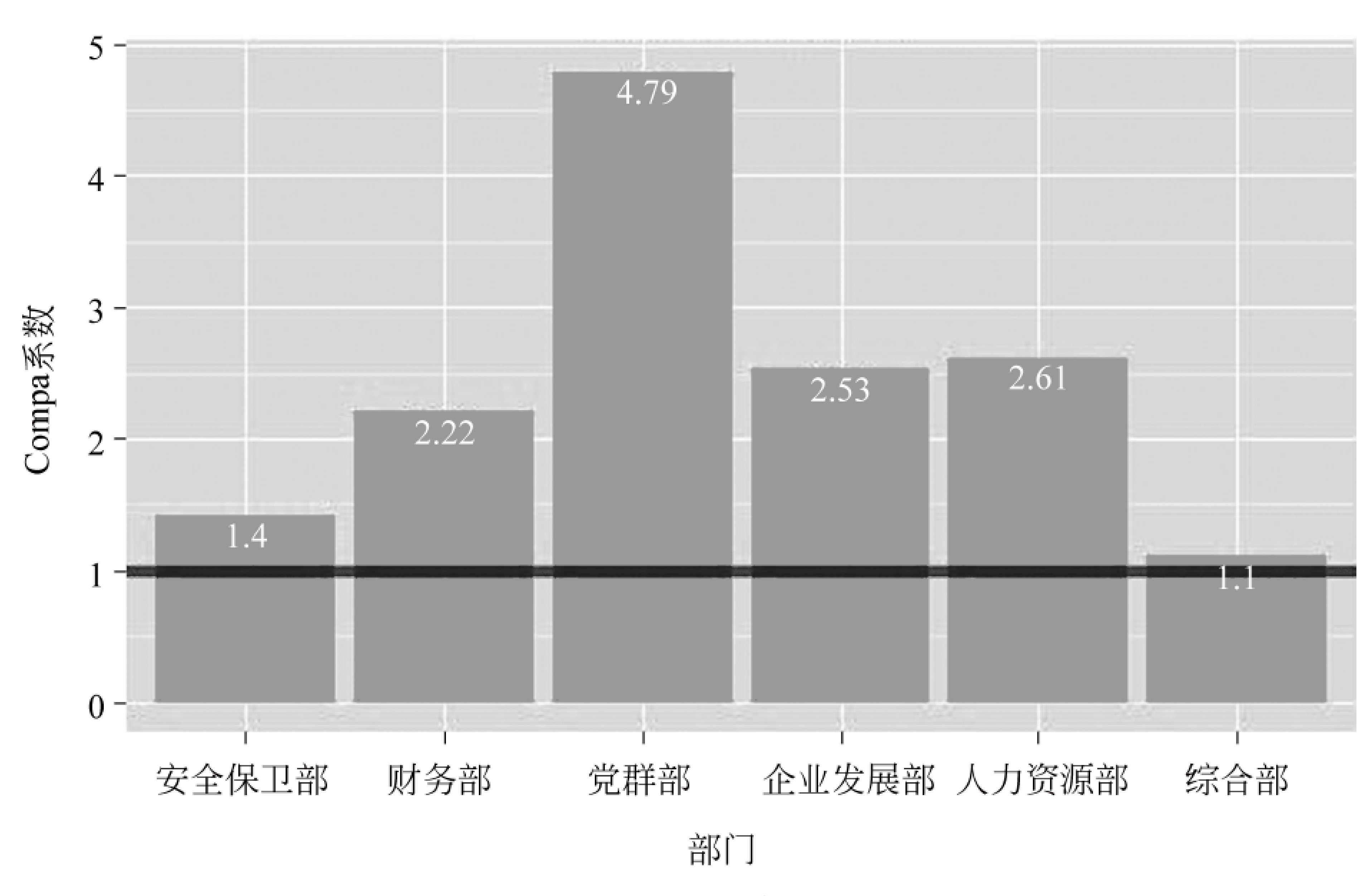


图 5-8　谦多顺公司职能部门 Compa 系数

R 语句如下：

```
library(ggplot2)
d<-read.csv("第五章/薪酬分析.csv")
d<-d[d$部门类型=="职能部门",]
d$部门<-droplevels(d$部门)
d.中位数<-median(d$应发工资)
d.平均数<-tapply(d$应发工资,d$部门,mean)
d.compa<-data.frame(Compa=d.平均数/d.中位数)
d.compa$部门 <-rownames(d.compa)
g<-ggplot(d.compa)
g+geom_bar(aes(部门,Compa),stat="identity",fill="red",
alpha=0.7)+geom_hline(size=2,yintercept=1,colour="blue",
alpha=0.7)+labs(title="职能部门 Compa 系数",x="",y="Compa")+geom
_text(aes(部门,Compa,label=round(Compa,2)),vjust=1,colour=
"white")
```

哇，看来党群部的薪酬水平很高啊。

Miss 陈：这可能和党群部人数少、岗位等级高有关系。总体来看，财务部、人力部、企业发展部等部门的薪酬水平是偏高的，安保部和综合部

处于均衡位置，这与我们对部门的价值评估结果是基本一致的。

小姚：那么，如果用公司薪酬和行业薪酬来计算薪资均衡指标，就可以分析公司薪酬在行业中的竞争力，是这样吗？

Miss 陈：是的。将公司的薪酬水平同行业薪酬中位数进行比较，计算薪资均衡指标，结果等于 1.0，说明公司的薪酬水平与行业是匹配的，即公司薪酬水平上涨幅度与通货膨胀水平相当；如果超过 1.0，则说明公司的薪酬水平领先于行业薪酬；如果小于 1.0，则说明公司的薪酬水平落后于行业薪酬。基于此，可以通过薪资均衡指标来判断企业的薪酬体系是否达到了人力资源管理的目标，是否匹配公司发展战略和发展阶段。

小姚：那么对公司内部的岗位而言，是否可以通过某岗位薪资，结合该岗位市场薪酬数据计算薪资均衡指标，来分析该岗位薪酬在市场中的竞争力呢？

Miss 陈：当然可以。我们可以用薪资均衡指标来分析公司各岗位薪资在社会行业中的相对地位，反过来也可以根据薪资均衡指标来调节这些岗位的薪酬水平。

小姚：太好了，这解决了长期困扰我们的问题。由于没有对标，没有计算薪资均衡指标这类数据，我们一直拿不准给员工的薪酬是高了还是低了。即使有了外部的薪酬数据，该如何对标，是对平均数还是中位数，也比较困惑。但是现在我可以计算各个岗位在行业中的薪资均衡指标，分析各个岗位的薪酬水平在行业中的地位，做出合理分析和评估。若某个岗位薪酬异常，就可以结合公司薪酬战略，调整和优化岗位的薪酬。

Miss 陈：是的。不过我们现在缺少行业薪酬数据，暂时还不能进行这样的分析。这类数据有公开的，也有商业的。公开数据有国家统计局、地方政府人力资源和社会保障部门发布的薪酬数据，但这类数据通常是用平均数计算得来的，很少见到用中位数。商业薪酬数据是通过市场调研公司、人力资源管理咨询公司、人才市场等单位定期收集整理而成，作

为产品销售。这类数据比较全面、详细，使用了平均数和中位数等数据，而且行业分类也较细，不过需要花钱购买。

5.4.4　用公平感计量模型分析员工对薪资的公平感

Miss 陈：最后我们试试用薪酬公平感计量模型分析员工的薪酬公平感。这个方法比前面的方法都要复杂一些，工作量会大一些。首先要收集数据，这个环节需要设计一份关于测评要素的等级评定表，采用 360°评价法开展问卷调查，通过上级、同级、下级评分，计算评价要素评分均值，才能代入公平感比率公式进行计算。

小姚：为了学习如何应用薪酬公平感计量模型，我先模拟一次评分吧。要设计评分表并开展调查得花不少时间，咱们暂时省去评价表设计和 360°问卷调查的环节，直接模拟评分结果吧。经理，您看行吗？

Miss 陈：可以。

小姚：太好了，我的模拟评分结果见表 5-4。

表 5-4　薪酬公平感计量模型模拟评分表

要素 类型	技能	知识	健康	任务	努力	现金薪酬	福利	培训	晋升机会
投入量	90	90	95	70	85				
回报量	70	70				80	80	40	50

将上面的评分数据代入薪酬公平感计量模型的计算公式，结果如下：

$$E_{公平感比率}=\frac{Q_{现金薪酬+福利+培训+晋升机会+知识+技能}}{I_{知识+技能+健康+任务+努力}}=1.01$$

从结果来看，我的薪酬公平感恰好在均衡状态，也就是说对感觉薪酬公平，哈哈。

Miss 陈：是的。不过从具体得分上看，你在技能、知识、健康、努力上的得分较高，这说明你是公司需要的人才，但获得的现金薪酬、福利维度

上得分一般，培训、晋升机会更少，工作中获得的知识和技能也有不足。虽然有薪资公平感，但应加强培训锻炼，适当增加工作任务，提供更多学习机会，促进你的快速成长，在此基础上还可提供晋升机会。

小姚：经理，您说得太好了，啥时候能有晋升的机会啊！

Miss 陈：加强学习吧，相信不久你就会有机会的。

第 6 章

员工综合能力评估

导语：企业在评选优秀人才时，需要对员工的综合能力进行评估，以区分优劣。但在实际评估时，往往带有较强的主观因素，导致出现误差，有误选、漏选的现象，没有选拔出真正优秀的人才。本章介绍如何使用综合评价法，将反映员工综合能力的各种评价指标进行量化，通过量化评估来选拔优秀人才。

6.1 需求描述

小曾：经理，根据今年的工作计划，本月就要启动公司优秀人才评选工作了。

Miss 陈：请按照工作计划组织优秀人才评选吧。

小曾：好的。不过根据去年的评选情况，今年我们想优化评选方法，主要是进一步提高优秀人才选拔的公平性和客观性。

往年我们评选优秀人才，主要是由各单位上报，我们组织专家评审小组进行评选。但在评选过程中发现评选的主观性很强，常会发生不太优秀的员工进入了优秀人才队伍，业绩不错的员工却被漏选了，效果不太理想。我听到一些单位和员工反映评选过程欠缺公平性。

Miss 陈：看来我们需要优化评选的方法，减少主观评价的影响以及因此造成的误差，选拔真正优秀的、综合素质能力高的员工，提高评选的公平性和客观性。

小曾：是啊。不过要怎么优化评选方法，减少评选过程中的主观性，提高选拔的公平性呢？这方面我还没有什么思路。

Miss 陈：可以使用综合评价法，结合目标优化矩阵、标准分等统计方法，将员工的各项能力要素进行量化，根据量化的结果来评定优劣。这些方法可以将主观判断转化为客观量化的数据，从而提高公平性。

6.2 分析方法

小曾：什么是综合评价法呢？

Miss 陈：解释综合评价法之前，我们先谈谈绩效考核吧。

小曾：和绩效考核有关系吗？

Miss 陈：是的，有些关系。现在说说你的月度绩效考核情况吧。

小曾：我每月的绩效考核是按照关键绩效指标(KPI)进行考核的，绩效考核表见表 6-1。

表 6-1　小曾月度关键绩效指标考核表

序号	工作类别	KPI	工作要求	分值/权重	评分方式
1	月度培训计划管理	月度培训计划编制	每月 28 日以前将各部门下月培训计划反馈给各个部门，做好培训实施前沟通、提醒工作	6	晚发一天/部门扣 1 分 漏发一个/部门扣 1 分
		月度培训计划变更	培训部门变更培训项目需经培训主管同意，培训主管于 30(31)日前将新培训计划反馈给培训部门	6	晚发一天/部门扣 1 分 漏发一个/部门扣 1 分
		公司临时性培训安排	公司领导下达培训指令后半个工作日内制订出培训方案，报领导批准后立即下发	5	无培训方案不得分 方案通过不下发不得分
2	培训管理	培训提醒	每项培训开展前 3 天提醒培训部门准备培训，提交培训教程和培训试卷，培训开展前 1 天咨询培训准备进展	5	少提醒一次扣 1 分 晚提醒一天扣 1 分
		培训材料管理	根据月度培训计划，检查培训部门培训材料准备情况	5	无检查 1 次扣 1 分
		培训协助	协助培训部门做好培训开展的相关工作，准备培训场所、培训设备以及培训材料等	6	影响培训正常开展 1 次扣 1 分 培训部门投诉一次扣 2 分
		培训抽查	每月抽查次数不得低于月度培训项目总数的 1/4	6	少一次扣 1 分
		培训效果评估	对每个培训项目培训的结果要检查、落实到位	5	少一次扣 1 分

续表

序号	工作类别	KPI	工作要求	分值/权重	评分方式
2	培训管理	异常报告	对培训过程中发生的异常问题要及时上报	5	有异常不报告一次扣2分
		培训总结及合理化建议	每月5日前将上月度培训总结上交	6	晚一天扣1分 无总结不得分
		培训档案管理	对每个培训项目要建立档案，形成培训档案库	6	少一份扣1分
3	课下培训资料管理	手指口述抽查、学习卡抽查	每个月对公司每个部门、分厂、工段至少抽查一个人做手指口述、学习卡抽查	4	少1人记录，扣0.5分
		导师带徒培训记录、培训笔记	每个月对公司每个部门、分厂、工段至少抽查一个人做培训记录、培训笔记抽查	4	少1人记录，扣0.5分
4	劳动纪律	考勤纪律	当月无迟到、早退，旷工、请假等	10	迟到、早退一次扣1分，旷工扣5分/天 请假一次扣1分(不满勤情况下)
		工作纪律	文明办公，团结同事，积极参与集体活动	5	与他人发生争执一次扣2分 不参加集体活动一次扣1分 不团结同事一次扣1分
5	个人提升	完善工作和学习技能	针对培训做出有效改进，每月接受一次专业培训	6	少一样扣2分
6	其他	领导交办的其他事项	爱厂如家，服从公司工作安排	10	视情节轻重

Miss 陈：那么你每月的绩效考核结果如何计算呢？

小曾：首先您根据我当月的工作表现，对考核表中的各个指标评分，然后将各项评分加权后计算总分，作为我的月度绩效考核结果。

Miss 陈：我们来数一下你的月度考核指标。嗯，你的月度绩效考核指标一共有 17 个，每个指标都评分的话，就会有 17 个分数，但是考核结果最终只是 1 个分数。重点在这里，你注意到了吗？这个最终的考核结果分数是不是综合了前面 17 个指标分数。

小曾：有些明白了，您的意思是月度绩效考核其实就是一种综合评价，是吗？

Miss 陈：是的。用绩效考核的例子来讲就容易理解了。考核表中的每个指标都是一个变量，考核结果是综合了每个指标的计算结果，综合反映了你的月度绩效表现。这种将多个变量转换为一个综合变量的分析方法，就是综合评价法，其核心是将多个指标转化为一个能够反映综合情况的指标，从而进行分析评价。比如，要衡量国家经济实力、地区社会发展水平、企业经济效益等，涉及很多因素，将这些因素综合成一个指标，就要运用综合评价法。

小曾：这么说来，我最近在网络上看到的最具幸福感城市排名、最适宜旅游城市排名等，都有具体的分数，这种排名分数就是用的综合评价法吧？

Miss 陈：是的。我们来看看综合评价法的分析步骤吧，如图 6-1 所示。

小曾：看上去综合评价法的分析步骤很清晰，也容易理解。但我对于其中一些步骤不知道如何操作，比如步骤 1，该如何确定指标体系呢？步骤 2，要如何对数据进行标准化处理呢？还有步骤 3，怎么合理地确定指标的权重呢？

Miss 陈：这些问题提得很好，下面我们就按分析步骤看看如何运用综合评价法。

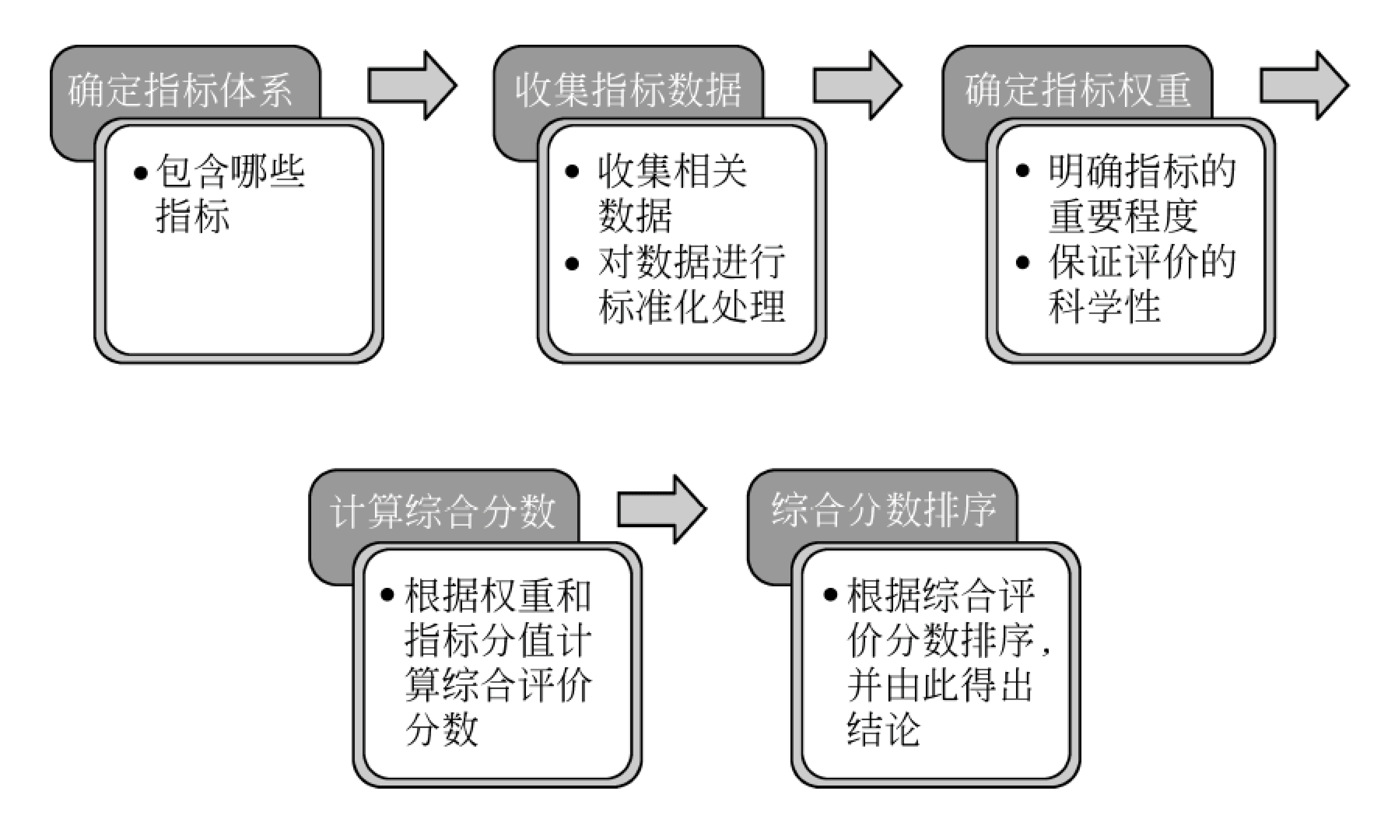


图 6-1　综合评价法的分析步骤

6.3　分析过程

6.3.1　确定指标体系

Miss 陈：回过来看需求，看如何在优秀人才评选中应用综合评价法。

小曾：好的，要怎么入手呢？

Miss 陈：根据综合评价法的分析步骤，首先我们要确定指标体系。

小曾：明白，就是要确定用哪些指标来评价员工的综合能力。但是要如何确定用哪些指标呢？

Miss 陈：确定指标体系的方法有不少呢，比如问卷调查法、专家访谈法、德尔菲法、聚类分析法和主成分分析法等。前三种是比较常见的分析方法，很多管理咨询公司都很擅长用这些方法开展咨询项目，姑且称其为咨询类方法；后两种属于统计学方法，要应用统计学知识，要收集数据并用专业软件进行统计分析，才能计算出指标体系，可称之为统计类方

法。如图 6-2 所示。

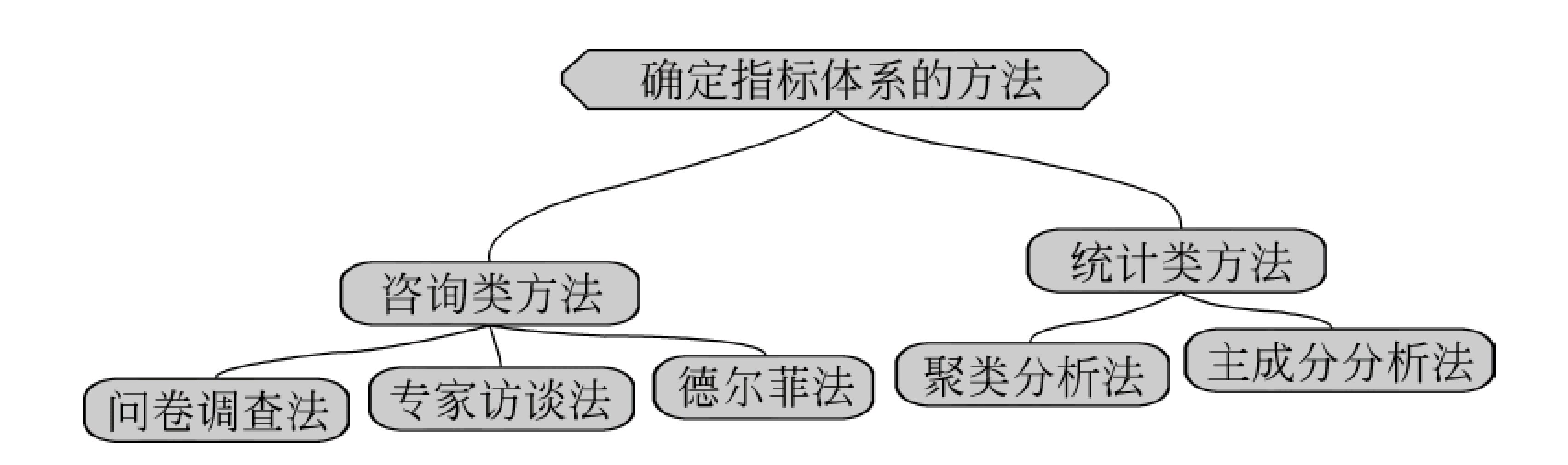


图 6-2　确定指标体系的方法

小曾：我们用过问卷调查法和专家访谈法。

问卷调查法是通过发放问卷来收集数据。这个方法在工作中经常使用，比如员工满意度调查就是一种问卷调查法。

专家访谈法是通过访问专业人士，根据访谈内容来确定指标体系。我们制定绩效考核指标的时候，也用到了这种方法。

德尔菲法虽然用得比较少，但我学习过这种方法。德尔菲法是采用匿名方式征询专家小组成员的意见，经过几轮征询，使意见趋于集中，最后做出分析结论的方法。

聚类分析法和主成分分析法我就不太清楚了。

Miss 陈：统计类方法比较少见，用得不多。由于涉及统计学知识，讲述会花很多时间，我先简单介绍一下。

(1) 聚类分析法：是根据“物以类聚”的道理，对数据或变量进行分类的一种多元统计分析方法。

(2) 主成分分析法：是将多个变量通过线性变换以选出较少个重要变量的一种多元统计分析方法。

如果我们预先收集的指标很多，有几十甚至上百个，多到难以判断哪些重要、哪些不重要，不知如何筛选时，就可以使用聚类分析法和主成分分析法。这类方法能够将多个指标进行降维(减少指标数量)，简化为几

个综合的指标。如果用统计类方法，除了收集指标外，还需要收集与指标相关的数据，有了数据才能进行统计分析。这两种方法都是多元统计分析的方法，涉及统计学知识，有兴趣的话你可以查查相关的资料。

小曾：统计类方法方法感觉很科学，回头查找资料学习一下。

Miss 陈：之前我们通过问卷调查法和专家访谈法，分析并确定了员工能力评价的指标，如图 6-3 所示。

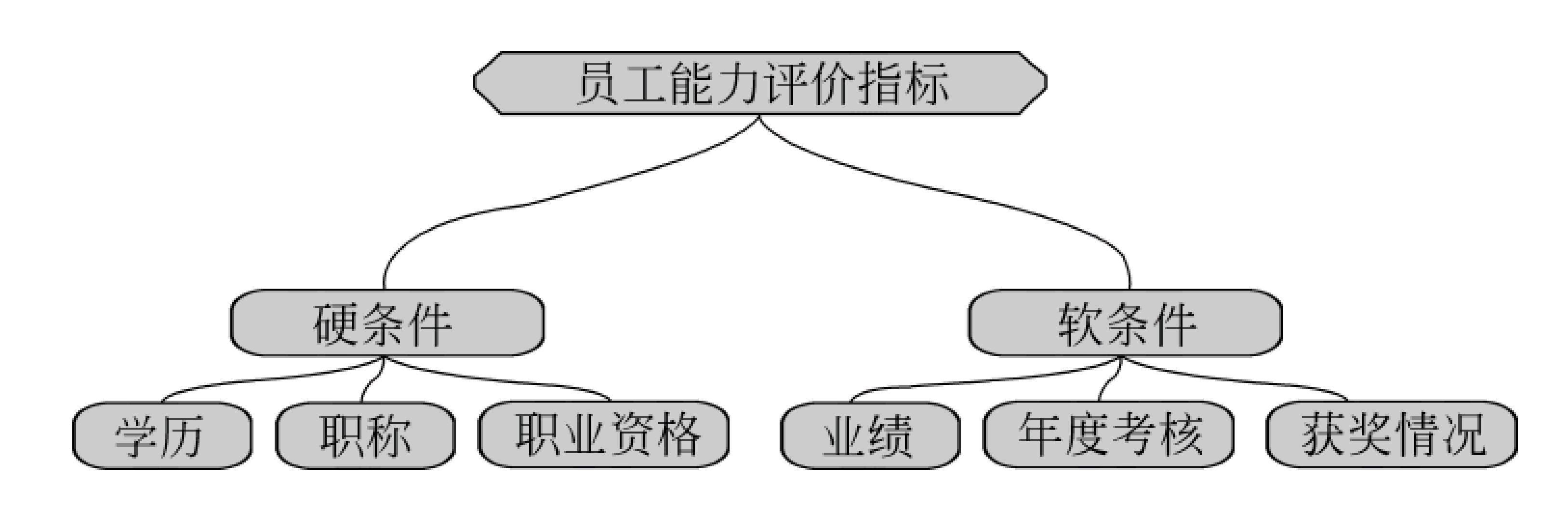


图 6-3 员工能力评价指标

虽然这个评价指标体系简单了些，许多指标都没有包括进去，比如发表论文、申请专利、培训授课、制度编写、流程改进、工作方法创新等，但作为讨论综合评价法的素材还是可以接受的。我们就用这套指标体系来研究如何通过综合评价法进行员工综合能力评估吧。

小曾：好的。原来还有这么多指标没有包括进去啊，以后我会根据您的指导进一步完善指标体系。

Miss 陈：我们现将指标体系转换为如下公式，再继续后面的内容。

$$Q=\text{学历}+\text{职称}+\text{职业资格}+\text{业绩}+\text{年度考核}+\text{获奖情况}$$

6.3.2 收集指标数据

小曾：按照分析步骤，接下来就要收集数据了。

Miss 陈：是的。由于今年的选拔还没开始，你可以把去年的数据拿出来试着分析。

小曾：去年的数据在我的电脑里，马上就可以调出来。好了，去年总共有 171 名候选人，基本情况见表 6-2。

表 6-2　优秀人才候选人基本数据示例

序号	姓名	所在单位	学历	学位	职称	职业资格	年度考核（2014 年）	年度考核（2015 年）	年度考核（2016 年）	所获奖励	标志性工作业绩
1	蔡海珠	E 分公司	本科	管理学士		**********	优秀	良好	称职	1. 2002*　**“********(**)********* ”,**********。 2. 2003*　**“********(**)*********”,***********。 3. 2004*　**“********(**)********* ”。 4. 2007*　**“ *******************”,******。 5. 2008*　**“*********************”,******。	1. 2009*　*********,C*******C*******,**9*****C*****。****C**********。*************。***********。 2. 2010*　*********,******************,**14*********。*****************,********************,*************。***********。 3. 2011*　*********,*****************,**17***。*******************。**********************。 4. 2012* **********,*********************,*********、****,*************、*********、******。*****************。*******2012*****。 5. 2013*　*********,**4G*****,******。****4G***********。**LTE******。

续表

序号	姓名	所在单位	学历	学位	职称	职业资格	年度考核（2014 年）	年度考核（2015 年）	年度考核（2016 年）	所获奖励	标志性工作业绩
2	曹德胜	G 分公司	本科	学士	工程师	*********	良好	优秀	优秀	1. 2006*8*,**“2006***********”**********,*******************、********,******。 2. 2001******“****”。 3. 2002******“****”。 4. 2003******“****”。 5. 2004******“****”。	1. 2010*8**2011*4*;****：**************************;****：*****,********;*******：****,****3361**。 2. 2010*1**2010*6*;****：*******************;****：*****,********;*******：****,****2167**。 3. 2013.03—2013.04 ****：2013*******************,****200000*,**5000*;****：*******;*******：****。****5000*。 4. 2013.04—2013.05****：2013*******************,****140000*,**6700*。****：*******;*******：****,****6700*。

续表

序号	姓名	所在单位	学历	学位	职称	职业资格	年度考核（2014 年）	年度考核（2015 年）	年度考核（2016 年）	所获奖励	标志性工作业绩
3	曹浩	K分公司	硕士	工学硕士	讲师	**** ****	优秀	良好	优秀	1. 2008-10，*****“***”******(*****)******* ***。 2. 2009-10，*****“***”******(*****)******* ***。 3. 2008****************-**(**)**。 4. 2010****************-**(**)**。 5. 2013****************-**(**)**。	****： 1. 2010-09 ********“ ****” **。 2. 2011-05 ***********************。 3. 2011-06 ********“ ****” **。 4. 2011-7 Android********。 5. 2012-02 ****“**2G*3G”3G*******(*********)。 6. 2012-04 ************。 7. 2012-06 *****************(******)。 8. 2012-09*****Android********。 9. 2012-10 ************。 10. 2013-07 **************。
4	曹陆元	Q分公司	本科	学士	高级工程师	****	良好	良好	良好	1. 1995*4*，**“*******”**，*********，******。 2. 1996*1*，**“**********”**，********，******。 3. 1997*6*，***********“****”，********，******。	1. 2011. 06—2012. 05　****10 000********　**********　******，*****10 000*******，*************，**********。****10 000*40%******，*10 000*********，****10 000************。 2. 2011. 3—2011. 12　****(**)******IP**********　**********　******，*********，*****************，********，*****************。

续表

序号	姓名	所在单位	学历	学位	职称	职业资格	年度考核（2014年）	年度考核（2015年）	年度考核（2016年）	所获奖励	标志性工作业绩
…	……	……	……	……	……	……	……	……	……	……	……
171	曾诗宇	T分公司	本科	学士	助理经济师		良好	优秀	良好	1. 2014*6*,**“2013******”**,***************。 2. 2013*12*,**“******”**,************ **。 3. 2013*12*,**“*****”**,************ **。 4. 2012*12*,**“*******”** ************ **。 5. 2012*12*,**“*****”**,************ **。	1. 2013.01—2013.12 ****2013************(***1.8**),********,******,*****2013*********。 2. 2013.04—2013.08 ****2014—2016*********(******40**),********,******,*********。 3. 2013.07—2013.12 ****2014****************(******35**),********,******,*********。 4. 2012.01—2012.12 ****2012************(***3.6**),********,******,*****2012*********。 5. 2012.08—2012.10 ****PTN**********(******32**),********,******,*********。

6.3.3　确定指标权重

Miss 陈：有了数据，接下来我们就要确定指标权重。

小曾：经理，我想问一下，为什么要确定指标权重呢，每个指标都取相同权重不行吗？

Miss 陈：权重相同显得简单粗暴，因为每个指标对人才综合能力的影响程度是不同的。打个比方吧，你觉得影响一部电影票房的因素有哪些？

小曾：我经常看电影，知道影响电影票房的因素有导演、演员、制作成本、宣传、口碑、档期、排片等。

Miss 陈：那么这些因素对票房的影响程度都是相同的吗？

小曾：在电影上映前期，导演、演员、宣传等因素的影响大些，上映中后期则口碑、排片等因素的影响大些。这么说来，每个因素对票房的影响程度都是不同的。

Miss 陈：对，回到我们的问题上，影响人才综合能力评估的因素有学历、职称、职业资格、业绩、年度考核、获奖情况，其实这些因素对人才综合能力的影响也是不同的。我们用这些指标来衡量人才的综合能力时，也要考虑到这些情况，要尽量区分各个指标对目标的影响程度，这种影响程度反映到变量上就是权重。

小曾：明白了，但是怎么确定指标权重呢？

Miss 陈：确定指标权重的方法也有好几种，跟前面确定指标体系的方法类似，也可以用问卷调查法、专家访谈法、德尔菲法等咨询类方法，也可以用聚类分析、主成分分析等统计类方法。还可以用一种比较简单的量化统计方法，叫作目标优化矩阵。

小曾：什么是目标优化矩阵呢？

Miss 陈：目标优化矩阵就是把模糊思维简化为计算机的 1/0 式逻辑

思维，最后得出量化结果的分析方法。这种方法不仅量化准确，而且简单、方便、快捷，推荐使用。下面我们按步骤来看看应该如何使用这种方法。

1. 建立矩阵

根据指标体系建立可用于两两比较的交叉矩阵，见表 6-3。

表 6-3　交叉矩阵表（待比较）

员工能力评价指标	学历	职称	职业资格	业绩	年度考核	获奖情况
学历						
职称						
职业资格						
业绩						
年度考核						
获奖情况						

2. 对比评分

以矩阵行指标为主线，依次与竖向列指标进行对比，根据指标重要性的对比情况进行评分。比如第一行第一个指标是学历，那么依次用学历跟竖向列的职称、职业资格、业绩、年度考核、获奖情况进行比较，若学历重要，则记 1 分，若不重要，则记 0 分。

小曾：我试着评一下分数吧。好了，您看看这样评分是否正确，评分结果见表 6-4。

Miss 陈：很好，就是这样评分。你应该注意到了，相同指标之间不用评分，所以从左上到右下的对角线用灰色做了标记，以这条对角线为轴，上下两个部分对称单元格的评分是相反的。

表 6-4　交叉矩阵表(已评分)

员工能力评价指标	学历	职称	职业资格	业绩	年度考核	获奖情况
学历		1	1	0	0	0
职称	0		0	0	0	0
职业资格	0	1		0	0	0
业绩	1	1	1		1	1
年度考核	1	1	1	0		1
获奖情况	1	1	1	0	0	

小曾：明白，这样就可以只评上半部分单元格的分数，下半部分的分数可以计算出来，可避免在指标很多的时候出现混淆。

Miss 陈：是的，接下来我们看看如何运用这些评分。

3. 优化矩阵

在已经评分的矩阵最右边添加一列，用来计算指标合计分。将每一行分数求和，填入该列。整理之后，就可以看到各个指标的重要性得分了。具体见表 6-5。

表 6-5　交叉矩阵表(计算合计分)

员工能力评价指标	学历	职称	职业资格	业绩	年度考核	获奖情况	合计
学历		1	1	0	0	0	2
职称	0		0	0	0	0	0
职业资格	0	1		0	0	0	1
业绩	1	1	1		1	1	5
年度考核	1	1	1	0		1	4
获奖情况	1	1	1	0	0		3

小曾：有点儿不对劲啊，职称得分为 0，这是说职称完全不重要吗？

Miss 陈：你注意到了这点很不错，在运用目标优化矩阵进行评分的时候，的确会遇到评分为 0 的情况，这时候需对评分结果进行调整。职称相比其他几个要素而言重要性较低，但也是衡量员工综合能力一个不可或缺的因素。所以，我们可以主动调整职称的评分结果。由于职业资格评分为 1 已是最低，那么可以给职称赋值为 0.5。修正后的评分结果见表 6-6。

表 6-6　交叉矩阵表（修正后）

员工能力评价指标	学历	职称	职业资格	业绩	年度考核	获奖情况	合计
学历		1	1	0	0	0	2
职称	0		0	0	0	0	0.5
职业资格	0	1		0	0	0	1
业绩	1	1	1		1	1	5
年度考核	1	1	1	0		1	4
获奖情况	1	1	1	0	0		3

4. 计算权重

接下来就可以根据合计分数计算权重了。方法很简单，用该指标的合计分除以总合计分，用百分数表示即可。

小曾：我来算算，结果见表 6-7。

表 6-7　评价指标权重

员工能力评价指标	合计	权重(%)
学历	2	13
职称	0.5	3
职业资格	1	6
业绩	5	32

续表

员工能力评价指标	合计	权重(%)
年度考核	4	26
获奖情况	3	19
合计	15.5	100

Miss 陈：根据计算结果，我们更新一下员工综合能力评价公式。

$$Q=\text{学历}\times 13\%+\text{职称}\times 3\%+\text{职业资格}\times 6\%+\text{业绩}\times 32\%+\text{年度考核}\times 26\%+\text{获奖情况}\times 19\%$$

小曾：公式更新之后，看上去更科学了。

6.3.4　量化指标内容

小曾：经理，我们收集的数据里面没有分数啊，您看"学历、职称、职业资格、业绩、年度考核、获奖情况"这些指标的内容都是文字性的，怎么计算分数呢？

Miss 陈：我们可以将这些文字内容进行量化。

小曾：量化？

Miss 陈：是的。首先，你有没有注意到，有些指标是有等级的。比如学历，填报内容包括中专、大专、本科、硕士研究生、博士研究生，这些内容其实是有等级区分的，学历从低到高，有顺序。再比如职称，填报内容包括初级职称、中级职称、高级职称，也有顺序。类似的指标还有职业资格、年度考核等。

小曾：对啊，学历、职称、职业资格、年度考核这几个指标都有明确的等级区分。

Miss 陈：这类有明确等级区分的指标，可以给每个等级赋值，等级越高赋值越高，通过赋值就将文字转换为数字，实现了量化，如图 6-4 所示。

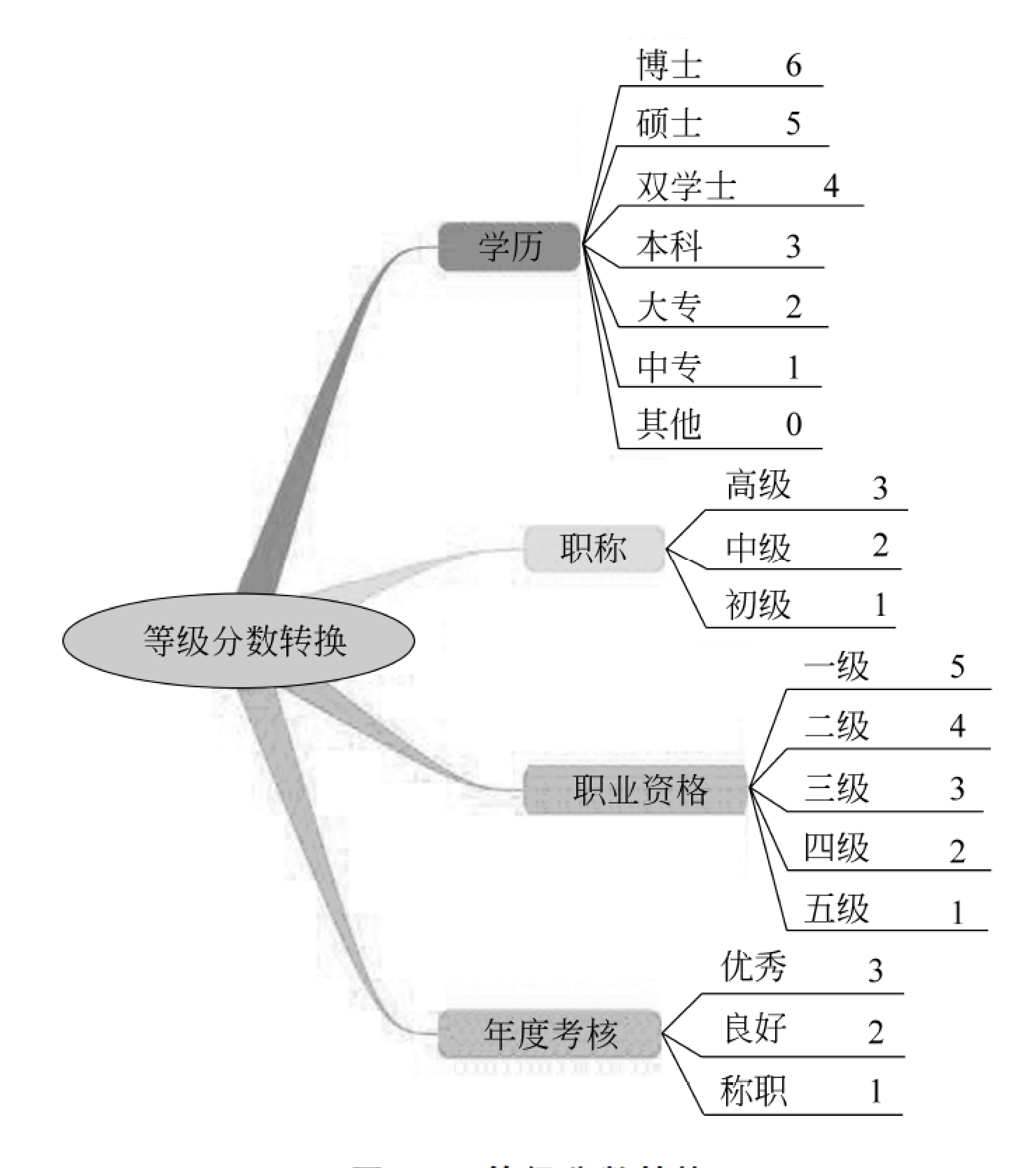


图 6-4 等级分数转换

小曾：哦，明白了。不过这种分数转换方法对有明确等级的指标有用，但是对于"业绩、获奖"情况这类没有明确等级的指标又该怎么办呢，而且这两个指标的内容很多，很难区分等级。

Miss 陈：这类指标要特殊处理。首先看获奖情况，其实奖励本身是有等级的，可以转换为分数，但还需要考虑到颁奖的单位也是有等级的。所以要结合这两个维度，综合制定评分表，才能将获奖情况指标合理转换为分数。评分表见表 6-8。

小曾：原来这样就可以把获奖情况的内容进行量化了。经理，如果一个人有多次获奖情况，分数可以累加吗？

Miss 陈：可以累加。

表 6-8　获奖情况组合评分表

授奖单位		特等奖	一等奖	二等奖	三等奖	其他
行政级别	公司级别					
国家级		7	6	5	4	3
省部级	总公司级	6	5	4	3	2
地市级	分公司级	5	4	3	2	1
县乡级		5	4	3	2	1

小曾：明白了，按照这个评分表来量化获奖情况指标，方便很多啊。

Miss 陈：最后是业绩指标，这个指标的量化比较复杂。我们可以采用等级评定法，制定一个分数等级。比如采用 10 级评分，10 分为业绩最好，0 分为无业绩，然后根据填报的业绩情况，由评分人主管评定一个等级分数，实现业绩指标的量化。等级设定表示如下。

0	1	2	3	4	5	6	7	8	9	10

小曾：这样评分的主观性比较强啊，由于每个人的评价尺度不一样，会导致有些人评分高，有些评分低。

Miss 陈：是的。为了降低主观性的影响，确保评分的客观性，可以组建评分专家小组进行群体评分。小组成员 3～5 人即可，由市场部、项目管理部的专家组成。每个成员分别对所有人进行业绩评分，然后取评分均值。

这种方法可以控制两个关键点，一是确保评分的专业性，因为成员由市场部和项目管理部的专家组成，可最大限度确保其评定业绩时的准确性；二是降低个人评分带来的喜好偏差，就是降低主观性对评分的影响，因为用了群体评分的平均分。

小曾：原来如此。我马上按照上述规则计算分数，业绩评分找市场部和项目管理部的几个专家进行评定。

（一周后）

小曾：经理，分数已经评定好了，结果见表 6-9。

表 6-9　人才综合能力评分表　　单位：分

序号	姓　名	所在单位	学历评分	职称评分	职业资格评分	年度考核评分	业绩评分	获奖情况评分
1	蔡海珠	E 分公司	1	3	0	6	96	36
2	曹德胜	G 分公司	1	2	3	8	98	26
3	曹浩	K 分公司	2	2	1	8	98	17
4	曹陆元	Q 分公司	1	3	1	6	96	6
5	曾诗宇	T 分公司	1	1	0	7	92	73
6	曾学	E 分公司	2	3	0	5	95	85
7	曾烨	E 分公司	2	2	8	4	95	34
8	陈东文	H 分公司	1	2	0	7	96	193
9	陈嘉莹	B 分公司	1	1	9	8	96.6	6
10	陈俊强	H 分公司	2	2	0	7	93	41
…	……	……	…	…	…	…	…	…

6.3.5　分数标准化

小曾：经理，现在可以计算总分了吗？

Miss 陈：还不行！你观察一下分数，感觉有什么不对劲的地方吗？

小曾：我看看，好像是有点儿不对劲，学历、职称、职业资格、年度考核这 4 个指标的分数都是 1 位数，但是业绩和获奖情况的分数有 2 位数甚至 3 位数的问题。

Miss 陈：观察得很仔细，这说明不同的指标量纲不同。

小曾：有什么影响吗？

Miss 陈：在量纲不同的情况下，计算结果会向量纲大的指标倾斜。

也就是说量纲大的指标会获得更大的权重，对结果的影响更大，量纲小的指标就显得微不足道了，对结果的影响有限。

小曾：那该如何消除量纲不同带来的影响呢？

Miss 陈：既然量纲不同，我们就让各个指标的量纲都变得相同吧。让量纲变得相同的方法也有好几种，我们这里采用将原始分数转换为标准分的做法（关于标准分的内容请参阅本书第 4 章）。

小曾：对啊，转换为标准分就能够统一量纲，您上次讲过呢。那我再计算一次，将原始分数转换为标准差为 10，均值为 100 的标准 *T* 分数，结果见表 6-10。

表 6-10　人才综合能力评分表（*T* 分数）

序号	姓　名	所在单位	*T* 学历评分	*T* 职称评分	*T* 职业资格评分	*T* 年度考核评分	*T* 业绩评分	*T* 获奖情况评分
1	蔡海珠	E 分公司	91.04	110.02	90.35	95.03	99.78	102.04
2	曹德胜	G 分公司	91.04	94.72	100.56	108.74	102.38	99.46
3	曹浩	K 分公司	109.73	94.72	93.75	108.74	102.38	97.13
4	曹陆元	Q 分公司	91.04	110.02	93.75	95.03	99.78	94.29
5	曾诗宇	T 分公司	91.04	79.42	90.35	101.88	94.58	111.61
6	曾学	E 分公司	109.73	110.02	90.35	88.18	98.48	114.71
7	曾烨	E 分公司	109.73	94.72	117.58	81.33	98.48	101.53
8	陈东文	H 分公司	91.04	94.72	90.35	101.88	99.78	142.63
9	陈嘉莹	B 分公司	91.04	79.42	120.98	108.74	100.56	94.29
10	陈俊强	H 分公司	109.73	94.72	90.35	101.88	95.88	103.34
…	……	……	…	…	…	…	…	…

转换标准分的 R 语句如下：

```
d<-read.csv("第六章/人才评价.csv")          #读取数据
d1<-scale(d[,16:21]) * 10+100              #转换为标准分
```

```
write.csv(d1,"第六章/标准分转换.csv")    #输出转换分数
```

Miss 陈：很好，经过标准分转换之后，各个指标的量纲就统一了，量纲不同的影响也消除了，这时就可以计算总分了。好了，现在再次修正一下前面的公式，这里用 T 表示标准分数。修正后的计算公式如下：

$$Q = T_{学历} \times 13\% + T_{职称} \times 3\% + T_{职业资格} \times 6\% + T_{业绩} \times 32\% + T_{年度考核} \times 26\% + T_{获奖情况} \times 19\%$$

6.3.6 综合分数排序

小曾：经理，经过权重分配、指标量化、标准分数转换后，后面的工作就轻松多了，只需要按权重计算总分就行了。

Miss 陈：是的，你来计算一下吧。

小曾：好的，计算完毕，结果见表 6-11。

表 6-11 综合评分总分

序号	姓　名	所在单位	总分(分)
1	蔡海珠	E 分公司	96.58
2	曹德胜	G 分公司	100.64
3	曹浩	K 分公司	102.22
4	曹陆元	Q 分公司	95.31
5	曾诗宇	T 分公司	97.60
6	曾学	E 分公司	99.22
7	曾烨	E 分公司	96.11
8	陈东文	H 分公司	105.62
9	陈嘉莹	B 分公司	99.84
10	陈俊强	H 分公司	99.33
…	……	……	…

Miss 陈：根据我们对人才综合能力评估的指标体系，经过对指标的权重计算，指标量化以及标准分转换，最后计算出总分。总分代表了人才综合能力，总分之间的差异反映了人才综合能力之间的差异。根据总分进行排序，分数由高到低，直观反映了人才综合能力的高低，由于进行了量化，人才综合能力之间的细微差别也能体现出来。计算结果能够给我们提供评价和选拔人才极为有用的、说服力强的信息。

小曾：是的。人才的综合能力被量化了，相互之间的比较也变得容易了。

6.4 结果应用

Miss 陈：不过在实际应用中，还需要注意以下两点。

1. 不唯分数

小曾：好不容易将员工的能力进行了量化，分数代表了员工综合能力的高低，我觉得很客观公正啊，为啥说要不唯分数呢？

Miss 陈：你回想一下量化过程，虽然我们把对指标的主观评价转换为了数字，但转换过程是会产生误差的。即使我们用了降低误差的技术，比如用群体评分取均值，但仍然不能避免误差的存在。如果两个人的综合评分很接近，那么将很难证明分数的差异不是因为误差造成的。

小曾：原来是这样。

Miss 陈：员工在填报个人资料的时候也可能出现误差。比如，有些员工实际是很优秀的，但因为工作忙碌，填报个人资料的时候内容过于简单，导致评分时分数偏低，脱离了实际情况。而有些员工工作比较闲，时间多，所以填报个人资料的时候就写了很多内容，这样评分时分数偏高的

可能性大。这些情况都有可能导致综合评分与实际情况不符。

小曾：这些情况我们很难控制，我们该怎么运用综合评价分数呢？

Miss 陈：建议把分数作为参考数据，再组织评审委员会进行评审，对个别有争议的员工，补充收集相关资料，对综合评价结果进行矫正。

小曾：明白了。

2. 员工能力评价指标体系还需要优化完善

Miss 陈：之前提到过，我们的员工能力评价指标体系还不够完善，有许多反映员工能力和价值的因素还没有纳入指标体系，这样会造成评价结果不够全面。

小曾：好的，接下来我会按照综合评价法的要求，进一步完善指标体系，更全面评估人才的综合能力，选拔出真正符合公司需求的优秀人才。

第 7 章

员工离职倾向分析

导语：员工离职会给企业带来损失，如果在员工入职前就能够预测该员工在一定时间段内的离职概率，将极大提高企业的招聘成功率，降低招聘成本。本章以招聘应届大学生为例，介绍如何用机器学习算法，根据招聘测评数据和学生信息建立预测模型，用模型预测新招聘大学生在入职一年内的离职倾向。

7.1 需求描述

小肖：经理，我们公司这几年招聘的应届大学生员工流失率比较高啊。我统计了一下相关数据，应届毕业生入职三年内的流失率都达到50％了。

Miss 陈：调查离职原因了吗？

小肖：是的。我对一些分公司做了调查，总结了一下大学生离职的原因，主要有以下几点。

(1) 认知偏差。这类大学生对公司以及工作的心理预期与现实环境落差较大，进入公司后发现跟之前憧憬的不一样，差别很大。这类大学生一般在入职后两三周就辞职了，还没过试用期呢。

(2) 适应性差。这类大学生不太适应周边环境，对气候、文化、语言等环境都不适应，出现适应困难症。一般离职后都回到老家去了，回到他们熟悉的环境中。

(3) 追求高薪资福利。这类大学生进公司只是找个落脚点，安定后马上找薪酬福利更高的公司，或者开网店自主创业。

(4) 内部管理原因。这类大学生或者与直接主管发生矛盾，或者对公司的管理制度不认可，或者对企业的文化不认可，由于公司内部管理原因而产生离职行为。

Miss 陈：这么看来应届大学生离职既有内部原因，也有外部原因，既有个人原因也有企业原因。既然找到了原因，我们就可以针对这些问题去改善。比如内部管理原因，如果是主管问题、制度问题、企业文化问题，可以分析具体问题，制订解决方案。

另外，我们也可以在招聘的时候，预测大学生在入职后的离职概率，提前判断他们是否会在一年内离职，作为招聘的参考依据，从而改善离职现象。

小肖：经理，您说在招聘的时候就预测大学生在入职后的离职概率？这不太可能吧。

Miss 陈：用数据分析的方法，这是可以实现的。不过要实现预测，我们需要收集一些历史数据，包括以下几方面。

(1) 近年招聘的应届毕业生招聘时的综合素质测评分数。

(2) 近年招聘的应届毕业生的个人基本资料。

(3) 应届毕业生入职后一年内的离职情况。

小肖：这些数据都保留了，不过，具体要收集几年的数据呢？

Miss 陈：年份越长越好。

小肖：好的，我马上去收集和整理数据。

7.2 案例分析

7.2.1 数据准备

小肖：经理，数据准备得差不多了。我收集了 2009—2012 年公司招聘的应届毕业生员工的数据，包括个人资料、综合测评分数、是否在一年内离职等，共有 1 459 人。具体包括以下信息：

```
[1] "序号"        "姓名"          "性别"
[4] "工作单位"    "工作单位类别"  "入职年份"
[7] "学历"        "毕业院校"      "专业"
[10] "职称"       "职业资格"      "是否党员"
```

```
[13] "言语理解"      "数学"              "逻辑"
[16] "常识"          "成就导向"          "抗压能力"
[19] "灵活性"        "影响性"            "支配性"
[22] "外向性"        "社交能力"          "心理感受性"
[25] "创新"          "敬业"              "情绪稳定性"
[28] "严谨性"        "完美主义倾向"      "录用时岗位级别"
[31] "是否有晋升"    "是否一年内离职"
```

其中个人资料包括“姓名、性别、工作单位、入职年份、学历、毕业院校、专业、职称、职业资格、是否党员、录用时岗位级别、是否有晋升”等。

综合测评分数是我们在招聘应届大学生时，两套测评问卷的分数。一套是胜任力测评，包括“抗压能力、外向性、社交能力、心理感受性、创新、敬业、情绪稳定性、严谨性、完美主义倾向”等指标；另一套是基本素质测评，包括“言语理解、数学、逻辑、常识、成就导向”等指标。

离职数据是入职一年后的离职情况，0 表示在职，1 表示已经离职。部分数据见表 7-1。

Miss 陈：很好，你的数据准备得很充分，接下来我们就可以进行分析了。

7.2.2 数据分析结果与解释

小肖：要怎么进行分析呢？

Miss 陈：作为比较，我们用两种算法来进行分析，分别是 Boosting 算法和随机森林算法。在使用这两个算法之前，我们先看看预测效果。

我根据你提供的数据，以“是否一年内离职”为因变量，其余的因素为自变量，用 Boosting 和随机森林算法建立了两个算法模型。有了这两个算法模型，就可以进行预测。

小肖：怎么预测呢？

表 7-1　应届毕业生基本情况及测评数据（简表）

序号	姓名	性别	工作单位	工作单位类别	入职年份	学历	毕业院校	专业	职称	职业资格	是否党员	言语理解	数学	逻辑	常识	成就导向	抗压能力	灵活性	是否一年内离职
1	梁××	男	N分公司	G类	2009	硕士	华南师范大学	人力资源管理	无	无	是	8	15.5	22.6	3.5	7.961 4	4.460 76	5.178 45	0
2	李××	女	N分公司	G类	2009	硕士	华南师范大学	应用心理学	无	无	否	9	11.5	17.8	4.2	6.010 03	4.460 76	3.541 26	0
3	傅××	男	N分公司	G类	2009	本科	广东技术师范学院	通信工程	助理工程师	无	是	10	9.5	12	4.9	5.619 76	6.862 73	5.724 19	0
4	叶××	女	N分公司	G类	2010	硕士	广东商学院	企业管理	无	无	是	9	12.5	15.9	4.9	7.571 13	4.460 76	6.269 92	0
5	韩××	女	N分公司	G类	2011	博士	暨南大学	光电检测	无	无	是	15	24.8	25.6	10.8	4.839 21	5.061 25	2.449 79	0
6	骆××	男	N分公司	G类	2011	硕士	华南师范大学	管理心理与人才测评	无	无	是	15	20.5	21.9	9.6	5.229 48	6.262 23	5.178 45	0
7	姚××	男	N分公司	G类	2011	硕士	华南师范大学	光电通信	无	无	是	10.5	22.5	24.5	10.8	5.375 83	7.112 93	4.632 72	0

续表

序号	姓名	性别	工作单位	工作单位类别	入职年份	学历	毕业院校	专业	职称	职业资格	是否党员	言语理解	数学	逻辑	常识	成就导向	抗压能力	灵活性	是否一年内离职
8	余××	男	N分公司	G类	2011	硕士	云南师范大学商学院	英语	无	无	是	13.5	22.5	17.9	10.8	7.571 13	5.961 99	4.086 99	0
9	姚××	男	N分公司	G类	2011	本科	华南理工大学	信息工程	无	无	否	15	22.5	16.5	9.6	5.229 48	5.361 5	7.361 38	0
10	蔡××	男	N分公司	G类	2011	本科	华南师范大学	通信工程	无	无	否	19.5	28.5	29	12	7.180 85	5.361 5	7.361 38	0
11	苏××	女	N分公司	G类	2011	本科	华南农业大学	英语（翻译）	无	无	否	18	24.5	37	12	3.619 6	3.560 02	8.452 85	0
12	林××	女	N分公司	G类	2011	本科	吉林大学珠海学院	英语专业	无	无	否	13.5	22.8	24.2	9.6	3.668 38	5.811 87	2.449 79	0
13	何××	女	N分公司	G类	2011	本科	广州体育学院	运动训练	无	无	否	12	10	20.8	9.6	1.717 01	2.659 28	5.724 19	0
14	孙××	女	N分公司	G类	2011	本科	广东财经大学（原广东商学院）	审计学	无	无	是	12	22.5	26.2	6	4.058 66	4.160 51	4.632 72	0

续表

序号	姓名	性别	工作单位	工作单位类别	入职年份	学历	毕业院校	专业	职称	职业资格	是否党员	言语理解	数学	逻辑	常识	成就导向	抗压能力	灵活性	是否一年内离职
15	卢××	男	J分公司	E类	2009	本科	澳门科技大学	财务学	无	无	否	9	10.5	24.8	7	4.448 93	2.959 53	6.269 92	0
16	高××	女	J分公司	E类	2011	本科	广东工业大学	会计学	无	会计从业资格证	否	13.5	10.8	24.8	7.2	4.839 21	6.462 4	4.086 99	0
17	梁××	男	J分公司	E类	2009	本科	合肥工业大学	电子商务	无	无	否	7	13	18.3	7	4.839 21	5.061 25	5.178 45	0
18	李××	男	J分公司	E类	2009	本科	华南农业大学	市场营销	无	无	是	7	14.5	20.1	4.2	4.839 21	7.463 22	6.269 92	0
19	杨××	男	C分公司	A类	2010	本科	重庆邮电大学	通信工程	无	无	否	7	4.5	11	3.5	5.229 48	3.560 02	5.178 45	1
20	张××	男	M分公司	G类	2010	本科	广东外语外贸大学	人力资源管理	无	无	否	6	7	5.7	2.8	4.839 21	4.460 76	4.632 72	1

Miss 陈：将新的应届毕业生数据，带入模型中进行运算，就可以得到预测结果。这类似于之前我们讲到的用回归分析模型预测员工人数。

小肖：现在还没有新的应届毕业生数据，可以在旧数据中随机找一个人的数据进行计算吗？

Miss 陈：可以。

小肖：那我随机找一个吧，数据见表 7-2。

Miss 陈：好，就用这名员工的数据。现在我将他的数据分别代入两个模型中进行“预测”，结果如下。

1. Boosting 模型

“预测”结果：

```
$formula
是否一年内离职~.

$votes
          [,1]      [,2]
[1,] 24.497 85  31.566 52

$prob
            [,1]       [,2]
[1,] 0.436 959 4   0.563 040 6

$class
[1] "离职"

$confusion
                Observed Class
Predicted Class 在职  离职
           离职     0     0
```

上面是 Boosting 算法模型给出的“预测”结果，结果显示：该员工不会离职的概率为 0.436 959 4，离职的概率为 0.563 040 6，离职概率>0.5，总体判断结果是“离职”。印证一下，实际上我们看到该员工的确在一年内离职了，计算结果符合实际情况。

表 7-2　某大学生员工基本情况及测评数据

序号	姓名	性别	工作单位	工作单位类别	入职年份	学历	毕业院校	专业	职称	职业资格	是否党员	言语理解	数学	逻辑	常识	成就导向	抗压能力	灵活性	是否一年内离职
19	杨××	男	C分公司	A类	2010	本科	重庆邮电大学	通信工程	无	无	否	7.00	4.50	11.00	3.50	5.23	3.56	5.18	1

2. 随机森林模型

"预测"结果：

```
在职 离职
0.22 0.78
attr(,"class")
[1] "matrix" "votes"

离职
Levels:在职 离职
```

上面是随机森林算法模型给出的"预测"结果，与 Boosting 模型"预测"的结果类似。结果显示，该员工一年内在职的概率为 0.22，离职的概率为 0.78，总体判断结果仍然是"离职"，计算结果符合实际情况。

Miss 陈：看到了吗，两个模型"预测"的结果一致，都与实际情况相符。

小肖：太令人惊奇了，竟然能够这么准确地预测出还未招聘的人员在入职后一年内的离职情况。如果实际招聘时能用其中一个模型进行预测的话，就能大大降低新员工在一年内的离职概率，这真是令人激动啊。

Miss 陈：是的。通过建立这种算法模型，在实际招聘工作中应用，就能提高我们招聘工作的精准度。我们还可以根据每年的实际情况，更新数据，不断优化模型，提高预测精度。

小肖：经理，为什么要用两种算法来进行分析呢？

Miss 陈：用两种算法可以相互印证，相互比较，看看哪种效果好。实际应用时，可选择准确率较高的算法。

小肖：您用的 Boosting 模型和随机森林模型这两种算法有什么特点呢？

Miss 陈：相比传统的回归分析、logistics 回归、决策树等算法，Boosting 模型和随机森林模型这两种算法具有更高的预测精度，更好的

自变量容许度，不需要自变量是数值型，也不需要检验自变量之间的多重共线性等问题，还能有效避免过度拟合的现象。

小肖：统计术语有点多啊，听不懂了。

Miss 陈：没关系，统计术语和原理不一定要搞得很清楚。咱们是坚持拿来主义，大概了解就行，能用就行。

小肖：好的。这两个模型给人的感觉挺抽象的，有具体形式吗？有没有计算公式？

Miss 陈：这两个模型的确是比较复杂和抽象，在刚才的计算过程中，模型存入了两个变量中，不太方便地展示模型的特征，如果你有兴趣可以看看 R 语言中的分析代码。

你有没有想过，刚才的模型中，我们用了很多自变量参加分析，这些自变量的重要性都是相同的吗？有没有混入无关紧要的自变量？哪些自变量才是重要的呢？

小肖：是啊，这么多自变量参与了分析，论重要性孰重孰轻呢？经理，是不是要像回归分析那样，对自变量进行筛选，去掉无关自变量呢？

Miss 陈：其实这两种算法都不需要像回归分析那样筛选变量。和传统的回归分析不同，这两种算法属于机器学习范畴，是分类算法中比较新的算法，具有很多优点。比如，这类算法不需要筛选自变量，自变量可以多达几千个，并且算法模型还能给出各个自变量的重要性。咱们分别来看看这两种算法模型计算的自变量重要程度。

1. 基于 Boosting 算法模型对各自变量重要性的分析

基于 Boosting 算法模型对各自变量重要性的分析如图 7-1 所示。

图中横条长短代表该变量对“是否一年内离职”这个因变量影响的重要程度，条形越长表示重要性越高。可以看出，专业、毕业院校、工作单位这三个变量是影响员工稳定性的重要因素。

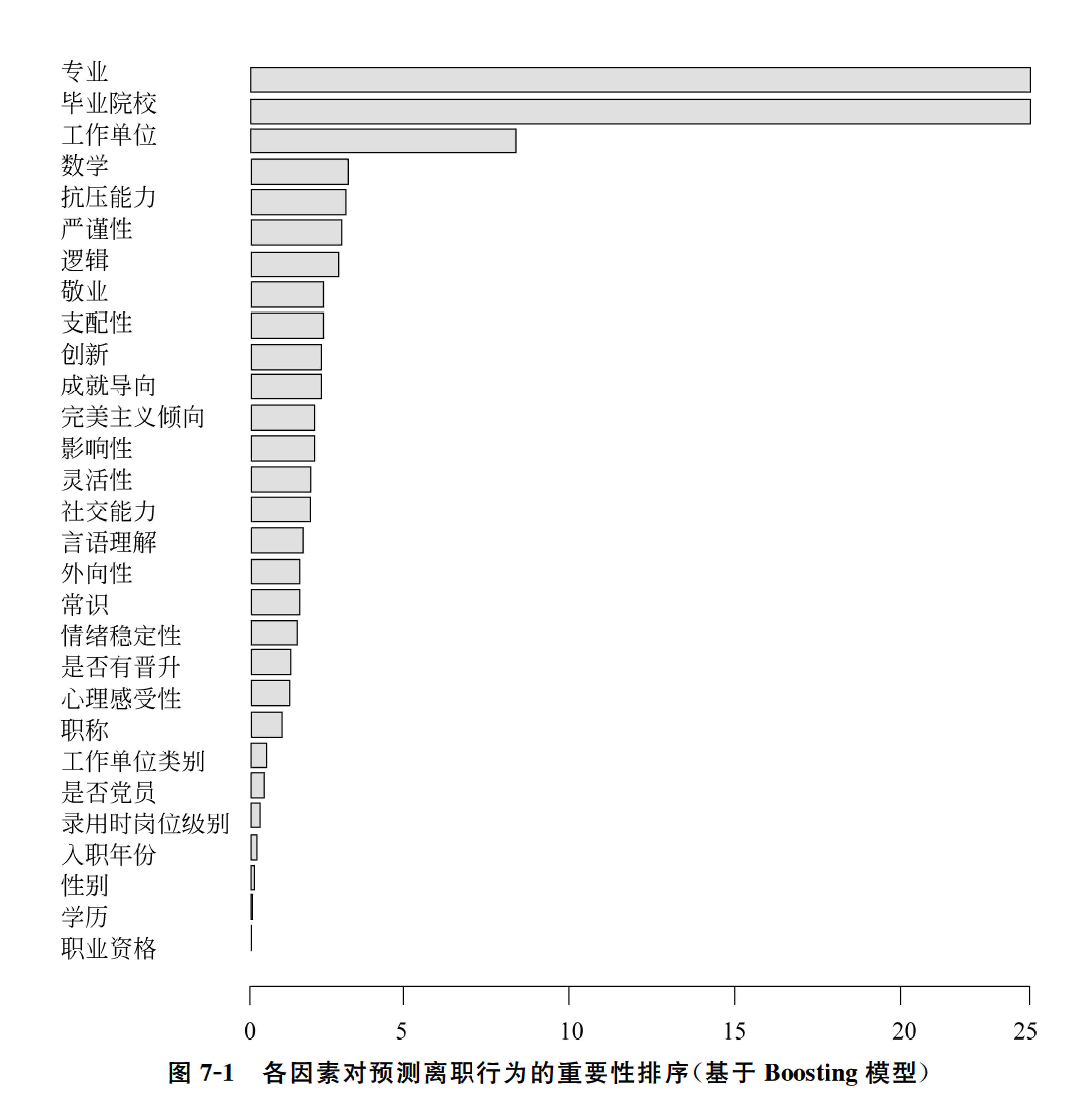


图 7-1 各因素对预测离职行为的重要性排序(基于 Boosting 模型)

2. 基于随机森林模型对各自变量重要性的分析

基于随机森林模型对各自变量重要性的分析如图 7-2 所示。

图 7-2 中,左边的图形是根据 MeanDecreaseAccuracy 来判断自变量的重要程度。MeanDecreaseAccuracy 是衡量指标,衡量把一个变量的取值变为随机数,随机森林模型预测准确性降低的程度。数字越大表示该变量的重要性越大。根据这个指标,是否晋升、工作单位、工作单位类别

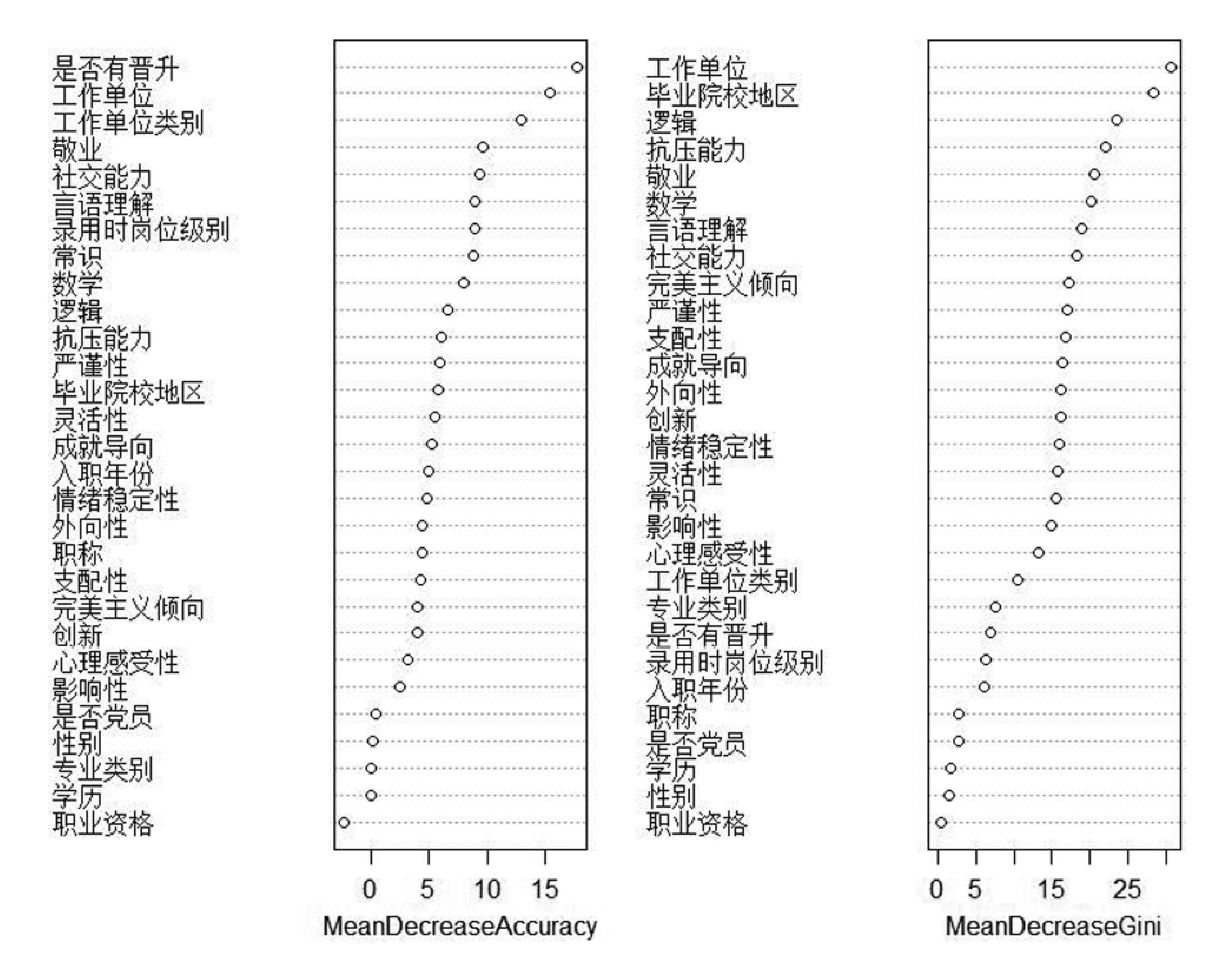


图 7-2　各因素对预测离职行为的重要性排序(基于随机森林模型)

三个变量是影响员工一年内稳定性的主要因素。

右边的图形是根据 MeanDecreaseGini 指数计算出的每个变量对分类树每个节点观测值异质性的影响程度,从而反映变量的重要性。该值越大表示该变量的重要性越大。根据这个指标,工作单位、毕业院校地区、逻辑分析能力等变量是影响员工一年内稳定性的主要因素。

小肖:咦,每种方法分析出来的变量重要性不大一样啊。

Miss 陈:是的,因为每种算法的原理和计算方式不同,判断变量重要性的策略不同,所以会出现这种情况。具体应用时,可以实际使用的算法为准判断变量的重要性。

小肖:总体来看,这真是令人惊讶的数据分析技术,它们竟然可以预

知未来。那么这两种算法工作的原理究竟是什么，要如何才能用它们开展分析工作呢？

Miss 陈：不急，接下来就说说分析方法和分析过程。

7.3 分析方法

7.3.1 Boosting 算法

Miss 陈：先说 Boosting 算法。该算法是为解决弱分类算法准确度不高的问题而提出的，从提出到现在，经历了好几个阶段。我们这里用的是当前普遍采用的 Adaboost（Adaptive Boosting 的简写）算法，可以翻译为自适应助推器算法。这种算法是一种迭代式的组合算法，目的是在不增加原始数据的情况下提高基础分类器的准确度，而我们在模型中用的基础分类器是决策树。

小肖：哦，就是说 Boosting 算法基于决策树，但采用了某种方法提高了预测精度。那决策树又是什么呢？

Miss 陈：决策树也是一种分类算法，是“在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成的图形很像一棵树的枝干，所以形象地称之为决策树。在机器学习中，决策树是一个预测模型，它代表的是对象属性与对象值之间的一种映射关系”。[①]

小肖：那为什么不直接用决策树算法来预测呢？

---

① 引用自百度百科。

Miss 陈：当然可以用决策树算法来预测。刚才你也提到了，决策树算法预测的误判率较高，准确率没有 Boosting 算法高。所以，我们在选择算法的时候，自然希望预测得越准确越好，是吧？

小肖：是的。

Miss 陈：Boosting 算法实际上是决策树的加强版本。决策树就是弱分类器，而通过 Boosting 算法，开始可能较弱（出错率高），然而随着迭代的进行，不断地通过自助法（Booststrap）加权再抽样，根据产生的新样本来改进分类器，每次迭代时都针对分类器对某些观测值的误判缺陷加以修正，每次迭代都根据这一轮产生的分类结果给出错误率，最终结果由各个阶段的分类器每轮错误率加权（权重是用来惩罚错误率高的分类器）投票产生，这就是所谓"自适应"的特点。

小肖：通过自身迭代提升准确度的思路真奇妙啊，感觉挺像人工智能，迭代加强后的预测效果有明显提升吗？

Miss 陈：是的，效果很好。Boosting 算法预测准确率相当高。比如，我们刚才对应届大学生离职行为的预测，以全部原始数据来做预测的话，误判率为 0，全部能准确预测。

小肖：真厉害啊！那这种算法有什么优、缺点呢？

Miss 陈：Boosting 算法的优点是预测准确率高、能够避免回归分析中的过度拟合现象，对自变量的类型和数量不挑剔，缺点是可能会被一些奇异点或者是离群点所影响。当然我们在实际应用的时候，还需要不断地完善和优化模型，添加更多的数据到模型中去优化模型，让模型更健壮。

小肖：经理，您多次提到分类器这个词，分类器又指什么呢？

Miss 陈：这是分类算法的一种叫法。比如"是否在一年内离职"这个变量，包括两种类别：在职、离职。我们在数据中用 0 和 1 来代表，这个变量就是一个分类变量。以这个分类变量为因变量进行回归分析、判别

分析等，这个过程就叫作类别分析，所用到的算法就是分类算法，也叫作分类器。

小肖：原来如此。

Miss 陈：分类算法的应用很广泛。一场足球比赛的结果是赢或输，一部电影的票房是高或低，一个顾客在超市中对某种产品买或不买，都涉及类别问题，可以用分类算法进行分析。

小肖：看来分类算法的应用真挺广泛。

7.3.2 随机森林算法

小肖：那随机森林算法也是一种分类器吗？也是通过某种方法将弱分类器变为强分类器的分类算法吗？

Miss 陈：是的。随机森林算法和 Boosting 算法有类似之处，都是通过某种方式增强分类效果。不同的是随机森林算法用了一种比较有意思的方法来进行自助抽样分类，这从算法的名称可以窥知一二。

小肖：是指“随机森林”这个名称吗？

Miss 陈：是的。随机森林也是以决策树作为基础分类器进行加强运算的，其中“随机”是指生成的决策树每个节点的变量仅仅在随机选出的少数变量中产生，每棵决策树所依据的数据都是随机的，连每个节点的产生都是随机性的。“森林”是指通过前述的随机方式生成了大量的决策树，这些决策树连起来就像是一片森林。这就是随机森林算法名称的由来。Boosting 算法在 R 语言中可默认生成 50 棵决策树，而随机森林则可默认生成 500 棵决策树，是不是名副其实的森林呢？

小肖：原来如此，随机森林这个称呼很形象。那么这个算法有什么优、缺点呢？

Miss 陈：优点和 Boosting 算法类似，分类精确度高、没有过度拟合的问题、对自变量类型容许度也高。对大数据，特别是自变量多的数据很

有效率，自变量甚至可以多达几千个，皆可以轻松应对，而且通过随机森林算法能找到重要变量。

小肖：看来随机森林是很优秀的算法。

Miss 陈：是的。Boosting 和随机森林都是机器学习类算法，发展历史不过二三十年。得益于不断进步的计算机技术，也得益于不断发展和普及的 R 语言，让我们普通人也能够使用这些先进的算法来解决实际管理工作中遇到的问题，帮助我们提升管理水平。

小肖：能用上这些算法真是太好了，那么实际的分析过程是怎样的呢？

Miss 陈：下面我们来看看分析过程。

7.4 分析过程

7.4.1 建模

Miss 陈：首先，咱们要根据现有的数据建立算法模型。以"是否一年内离职"为因变量，其余的维度为自变量，分别建立 Boosting 和随机森林的模型。

1. 建立 Boosting 模型

下面的 R 语句将读取数据，并用 Boosting 算法建立模型，存储到变量 m 中。

```
library(adabag) #Boosting 包
#读取数据
d<-read.csv("第七章/毕业生数据 1.csv")
d<-d[,3:32]
```

```
d[,"是否一年内离职"]<-factor(d[,"是否一年内离职"])
levels(d[,"是否一年内离职"])<-list(在职= 0,离职= 1)
#建立 Boosting 模型
set.seed(4410)
m<-Boosting(是否一年内离职~.,d)  #建立模型
```

2. 建立随机森林模型

随机森林算法对数据的要求要稍微严格些。对因子类的变量，也就是分类变量，要求不能超过 53 个类别，否则不能进行建模。所以在进行随机森林的建模时，需要将原始数据中类别超过 53 个的变量进行转换，这涉及“毕业院校”和“专业”这两个变量。我进行了归类转换，“毕业院校”按学校所处省份归类转换，“专业”按学科分类。

下面的 R 语句将读取数据，并建立随机森林模型，存储到变量 m_1 中。

```
library(randomForest) #随机森林包
#读取数据(对水平超过 53 的变量进行整理,归类降低水平数量,比如毕业院校)
d1<-read.csv("第七章/毕业生数据 2.csv")
d1<-d1[,3:32]
d1[,"是否一年内离职"]<-factor(d1[,"是否一年内离职"])
levels(d1[,"是否一年内离职"])<-list(在职=0,离职=1)
#建立随机森林模型
set.seed(101010)
m1<-randomForest(是否一年内离职~.,data=d1,proximity=TRUE,
importance=TRUE,na.rm=TRUE)  #建立模型
```

小肖：看上去建模过程简洁快速，读取数据之后立马就建好算法模型了，而且就用了一条语句。

Miss 陈：是的。其实许多算法都是原理复杂、解释复杂，理解起来困难，但是应用却是比较简单的。Boosting 和随机森林的建模过程只需一条语句、一个函数就完成了。不过在执行这条语句的时候，会花一点时间，时间多少取决于数据量大小，数据越大耗时会越多。比如，我们对离

职倾向的分析，有 30 个变量、1 459 条数据，执行建模语句消耗的时间大约要一分钟。

小肖：时间不算长啊。

Miss 陈：如果数据量很大，有上百万条数据的时候，计算时间就会很长了。还好我们在企业管理中所面对的数据量都不算大，不用处理大量的数据。

7.4.2　检验

Miss 陈：接下来我们检验一下两个模型的预测效果。

小肖：好的，要怎么检验呢？

Miss 陈：我们就拿原始数据来检验吧。将原始数据代入模型中进行模拟"预测"，将"预测"结果和实际结果进行比较。

1. Boosting 模型预测效果检验

Boosting 模型效果检验的 R 语句如下：

```
p<-predict(m,d)          #用原始数据进行预测
table(d$是否一年内离职,p$class) #查看预测结果与原始数据之间的差异情况
```

预测结果与实际情况的对比见表 7-3。

表 7-3　预测结果与实际情况的对比（基于 Boosting 模型）　　单位：人

实际＼预测	离职	在职
在职	0	1 217
离职	242	0

从表 7-3 可以看到，预测在职 1 217 人，实际在职 1 217 人，实际离职

242人，预测离职242人，误判率为0，预测效果相当好。

2. 随机森林模型预测效果检验

随机森林模型检验效果的R语句如下：

```
p1<-predict(m1,d1)  #用原始数据进行预测(直接给出分类结果)
table(d1$是否1年内离职,p1)  #查看预测结果与原始数据之间的差异情况
```

预测结果与实际情况的对比见表7-4。

表7-4　预测结果与实际情况的对比（基于随机森林模型）

单位：人

实际＼预测	离职	在职
在职	1 217	0
离职	0	242

从上表可以看到，随机森林模型预测的结果与实际的情况完全一致，误判率为0，与Boosting模型一样具有非常好的预测效果。

7.4.3　应用

小肖：经理，今年应届毕业生的招聘咱就用这算法来分析预测吧。

Miss陈：可以的。

小肖：实际应用的时候，是否将每个大学生的数据代入模型中，就可以预测入职后的离职概率了？

Miss陈：是的。但提醒一点，事无绝对，这类算法虽然有很高的预测精度，但仍然是建立在概率统计基础上的算法，仍然有误判概率。实际应用中应将其预测结果作为参考，而不应作为招聘标准，完全依赖计算结果。

小肖：好的，我们会以预测数据作为参考，结合招聘经验和其他信息，综合决策。

Miss 陈：关于 Boosting 和随机森林算法，背后有复杂的统计学原理，我们没有展开讲解。如果你感兴趣，可以找这方面的资料来学习。这里再次表达对 R 语言的喜爱和赞美，如果没有 R 语言，要使用上这类算法将会非常麻烦。

小肖：好的，谢谢经理。

第 8 章

员工辞职报告的情感分析

导语：离职面谈是员工离职管理的重要内容，如果在离职面谈时能够提前掌握员工离职前后的情感要素，将有助于提高离职面谈的成功率。本章介绍如何运用数据分析中比较少见的文本分析方法，从员工提交的辞职报告中挖掘情感信息，掌握员工离职时的情感线索。

8.1 需求描述

小肖：经理，向您汇报一件事。

Miss 陈：什么事？

小肖：最近有几个骨干员工离职。我跟他们进行了离职面谈，希望了解离职的原因，并尽量挽留。但他们都说得含糊其词，摸不清楚他们提出离职的真实原因，也没有挽留成功。

Miss 陈：如果员工主动辞职，那么在离职面谈时，员工通常会很谨慎。出于保护自己、顺利离职的心理倾向，不太愿意吐露真实想法。一般要等离职完成后一段时间，才有可能说出真正的原因。

小肖：那怎么办呢？这样离职面谈好像就没有必要了。

Miss 陈：离职面谈自然是很重要的，一次离职面谈就是一次管理咨询的过程。离职的时候是员工关系管理的脆弱阶段，如果工作做得好，将能挽留住员工，减少损失。否则极有可能使离职员工心怀不满，激化矛盾，甚至引起劳动纠纷。离职管理同时也影响着在职员工的情绪和心理。在员工离职管理诸多环节中，离职面谈的作用不可低估。

小肖：那应该怎么进行离职面谈才有效果呢？

Miss 陈：离职面谈的关键点是了解员工离职的原因。首先要与员工的所在部门沟通，了解部门经理对于员工离职的态度，确认挽留该员工的必要性，了解员工当前工作的进展情况，以及如果离职在什么时候交接工作比较合适，人力资源管理部门如何配合部门经理掌握离职进程的安排，等等。然后尽可能收集该员工的信息，包括近段时间的绩效考核情况、同事之间的口碑、劳动合同状况等，还需要查阅和分析员工的辞职报告。

小肖：经理，您说的大部分工作我们都会处理，但是辞职报告需要查阅和分析吗，有什么用意呢？

Miss 陈：辞职报告比较重要。首先，辞职报告能证明员工是主动辞职，在法律上有效力，这是其最重要的作用。其次，我们还可以从辞职报告中分析员工潜藏的情绪特征。

小肖：从辞职报告中分析员工潜藏的情绪特征，这个咱们能办到吗？

Miss 陈：可以的，我们可以用数据分析的方法来探索员工的情绪特征。

小肖：但是，分析员工的情绪特征有什么意义呢？

Miss 陈：主要是配合离职面谈使用，能够帮助我们较准确地把握员工的情绪特征，采取相关的应对措施。对主动离职的员工，离职面谈有两个目的，一是希望挽留核心员工，二是了解离职动机。通过离职面谈可以分析离职是否与企业管理或者政策有关，排除管理隐患，如果发现企业管理存在的问题，就要及时补漏，避免多米诺骨牌效应。

如果我们在离职面谈之前，已经了解和掌握了员工在离职时潜藏的情绪特征，那么在面谈的时候是不是更有主动权呢？是不是能找到与员工类似的情绪体验，更能理解员工的心理呢？如果与员工在心理层面拉近了距离，员工是不是更容易吐露真实想法呢？

小肖：这么说来，了解员工的情绪特征用处还挺大，但是怎样做才能从辞职报告中分析出员工的情绪特征呢？您说用数据分析的方法，但是辞职报告都是文字，怎么进行数据分析呢？

Miss 陈：虽然都是文字，但也可以用数据分析的方法，这种分析方法属于文本分析，我们要用到文本分析技术中的情感分析法。

小肖：文本分析？情感分析？听上去不太明白呢。

Miss 陈：请找一份员工的辞职报告来吧，我们模拟分析一次就清楚了。

小肖：好的。

8.1.1 数据准备

Miss 陈：你准备好员工的辞职报告了吗？

小肖：准备好了，下面是一位员工的辞职报告，原文[①]如下。

尊敬的××领导：

在递交这份辞呈时，我的心情十分矛盾。现在公司的发展需要大家竭尽全力，由于我状态不佳和一些个人原因的影响，无法为公司做出相应的贡献，因此请求允许离开。

从昨天晚上到今天，是继续坚持还是果断放弃？这个问题一直困扰着我，经过一天一夜的考虑，我还是选择放弃。曾经那么大的风浪都挺过来了，多么难过的坎也过去了，如今为什么就选择放弃呢？这个问题我也问过自己，回想这两年来，其实我比任何人都珍惜这份工作，我知道我不是最优秀的，但我是非常努力的。值得庆幸的是至少学生对我的责任心和努力付出是非常肯定的。

仅凭学生的几句话，我就知道我这两年的努力是值得的。

企业和领导对员工的肯定与鼓励、关心与爱护，就如同老师对学生的肯定与鼓励、关心与爱护，让学生在精神上取得莫大的鼓舞从而会更加自觉地努力学习；对于员工来说，也会发自内心的、心甘情愿地为企业做贡献，使企业得以更好的发展。其实我认为这是相互的，老师对学生不付出爱，学生当然也不会爱老师，少一点私心，多一点无私，真心真意为学生着想，学生才会真的爱老师，班级才会团结稳定。从 2006 年 9 月带 A 班和 B 班开始，我一直这么认为，也一直这样努力着，可是渐渐地，我发现因为

① 本辞职报告的内容纯属虚构。

各方面的因素这一点越来越难做到了,可能是我太单纯,可能是我不成熟,可能是我太理想化,我感到迷茫和困惑。做任何事我都很努力、很认真,我不想混,那样会让我良心不安,同时我胆小,做事之前,都要思来想去,瞻前顾后,害怕出现自己不能预料的后果,缺乏魄力,这也是我在工作上难以得到发展的致命缺点。不果断,是我另一个致命的缺点。

如今我的工作也真的走到了瓶颈处,因此我不得不离开。

也许此时提出辞呈会显得不合适,公司正处于快速发展的阶段,同事们都是斗志昂扬,壮志满怀,而我在这时候却因个人原因无法为公司分忧,实在是深感歉意。其实我也很不舍,舍不得那些既让人爱又让人恨的学生,舍不得相处了两年的同事,感谢向老师一直以来对我的关心和教导,也感谢肖老师平时对我的关心和帮助,在他们身上有很多值得我学习的地方,向老师超强的处事能力,肖老师圆滑的为人处世风格,这些都是我非常缺乏的。更加感谢公司给了我做老师的机会,感谢公司所有领导和同事对我的教诲和关心,这两年我也收获颇丰。

最后,我有一个请求,从去年周年庆到今年周年庆,我又工作了整整一年,虽然没有取得过人的成绩,但我是在勤勤恳恳地做事,我从来不会提出要求,这也是我的第三个缺点,这是我第一次也是最后一次要求,恳请领导在结算工资时将周年庆奖金和 2006 年 9 月至 2007 年 12 月所欠的社保费用连同工资一起结算给我,不胜感激!

我希望公司领导在百忙之中抽出时间商量一下工作交接问题。本人将于 2008 年 9 月 5 日离职,希望得到领导的准许!感谢诸位在我在公司期间给予我的信任和支持,并祝所有的同事和朋友在工作和活动中取得更大的成绩和收益!

此致

敬礼!

Miss 陈:好的,接下来我们看看应该如何进行分析。

8.1.2 分析结果与解释

Miss 陈：由于分析过程比较复杂，为便于理解，我们先看分析结果，再谈分析过程。

辞职报告的情感类别分析结果如图 8-1 所示。

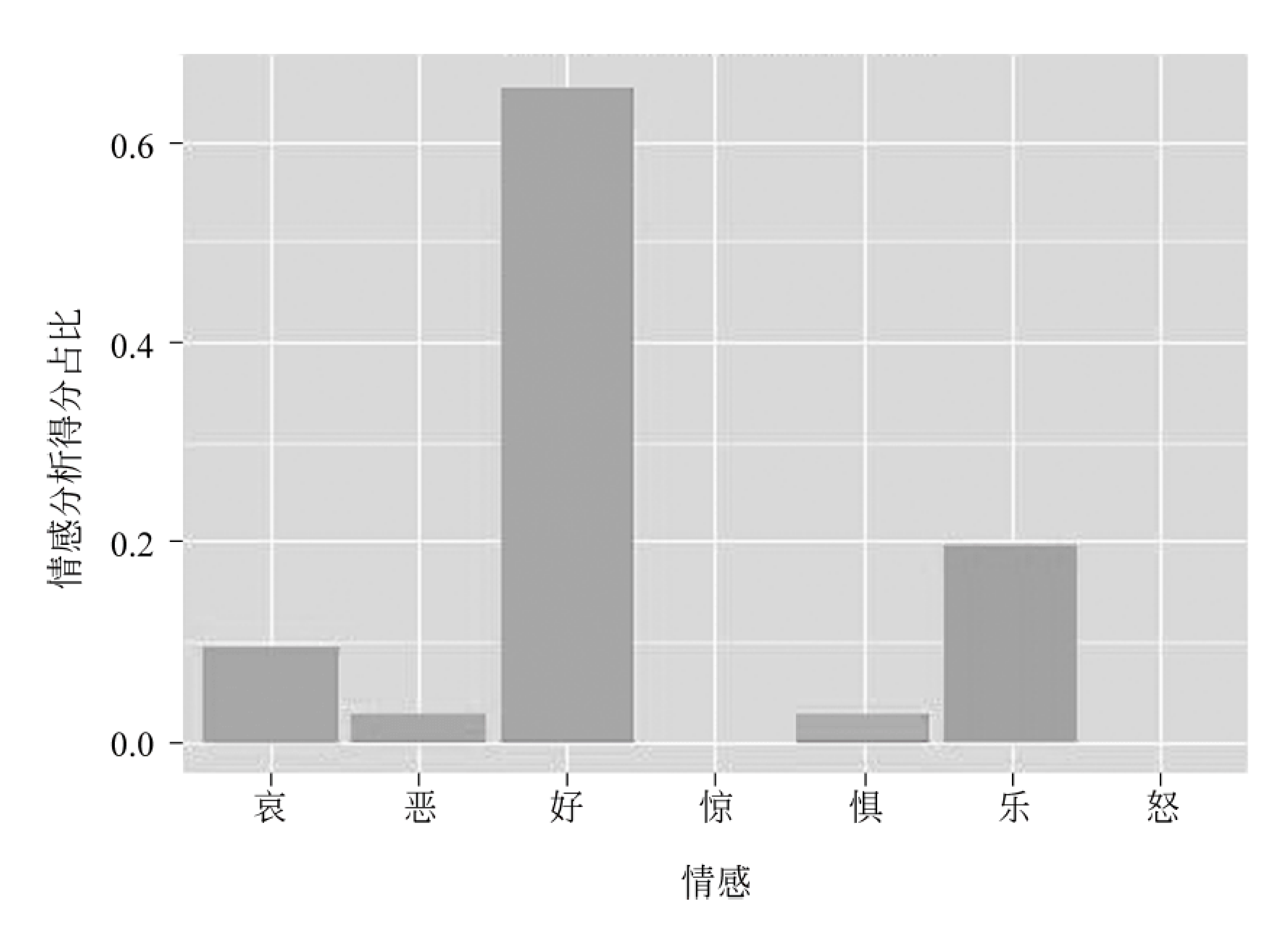


图 8-1 辞职报告的情感类别分析结果

辞职报告的情感极性分析结果如图 8-2 所示。

图 8-1、图 8-2 是辞职报告的两种情感分析结果，一是按情感类别分析；二是按情感极性分析。横坐标分别是情感类别和情感极性，纵坐标是各种情感类别和情感极性得分在总分中的占比。

小肖：图形显示的结果很直观，各类情感的重要性一目了然。

Miss 陈：是的。比如对这份辞职报告的情感类别分析，可以看到主要情感是“好”；其次是“乐”；再次是“哀”，且“好”是主导情绪。

“好”反映尊敬、赞扬、相信、喜爱的情绪，“乐”反映快乐、安心的情绪，

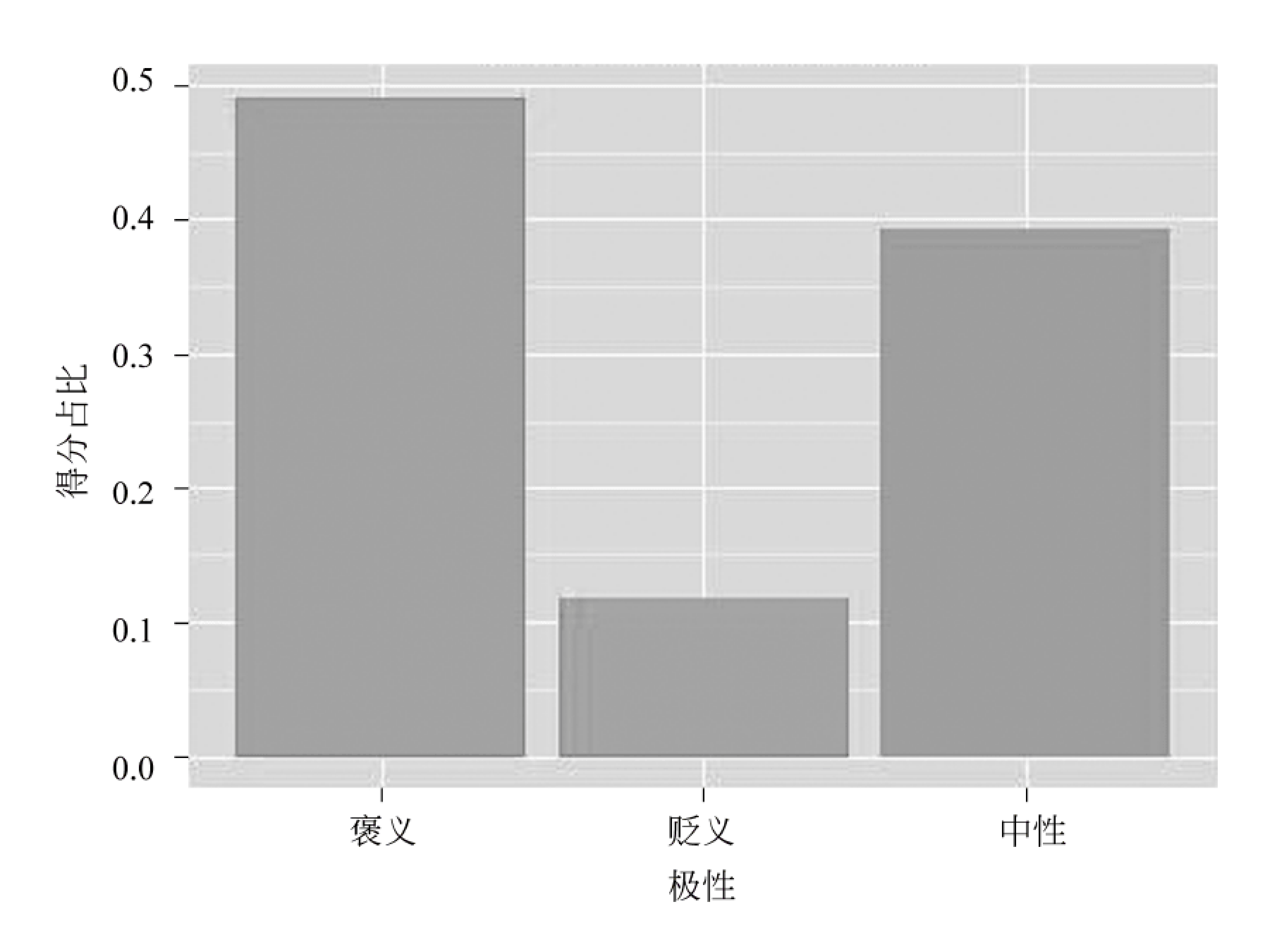


图 8-2　辞职报告的情感极性分析结果

这就是整篇辞职报告反映出的情绪特征。其中也有一点“哀”的情绪，哀反映悲伤、失望的情绪，在辞职报告中有这种情绪可以理解。还好在这篇辞职报告中这类情绪占比并不高。

从情感极性分析，可以看出辞职报告的主要情感极性是褒义；其次是中性；贬义情感虽有但占比不高，这也印证了情感类别分析的结果。

总体来看，这篇辞职报告以正面情绪为主，情绪相对积极、平稳，没有包含过多的负面情绪，辞职原因不太可能是对工作不满、与主管冲突而产生怨气所致。

小肖：真不可思议，您的分析和实际情况基本是吻合的。我了解过这名员工的情况，他本来工作挺好的，辞职原因是受同学的邀请而去创业。

Miss 陈：这些都是根据分析结果所进行的推论。不过员工为顺利辞职通常会倾向于隐藏情绪，所以即使负面情绪占比少，也要重点关注。

比如仔细阅读辞职报告，会发现员工提到拖欠奖金和社保费用的事情，这点要引起重视，很可能不是个别现象。情感分析可以为离职面谈提供分析素材，具体应用还要看实际情况。

小肖：明白了。那么您快讲讲如何进行情感分析吧，咱等不及要学习一下这方面的知识了！

Miss 陈：好的。

8.2 分析方法

8.2.1 文本内容的情感分析方法

Miss 陈：首先介绍一些基础知识。一篇文章反映了什么情感？褒义还是贬义？肯定还是否定？反映喜、怒、哀、乐、愁中的哪些情感特征？对这些问题的分析就是情感分析，或者叫情感倾向分析。有正常阅读能力的人，在看了一篇文章后也能够判断文章表达的情感。但这是主观评价，不够精确，不是量化数据。在对文章进行文本分析的时候，通常要将文本内容进行量化转换，才能够更加直观、精准地分析。

情感分析有两种方式，一种是情感极性分析，一种是情感类别分析。前者分析文章的总体情感态度，是“褒义”“贬义”还是“中性”，后者分析文章反映了哪种情感？具体来说有“乐”“好”“怒”“哀”“惧”“恶”“惊”等情感类别。

小肖：明白，原来情感极性分析就是分析文章的褒义、贬义等极端情绪，情感类别分析就是分析文章的情感类别。但是用什么方法进行分析呢？

Miss 陈：情感分析的方法有两类，一类是基于情感词典的方法，一

类是基于机器学习的方法。

小肖：什么是基于情感词典的方法呢？

Miss 陈：基于情感词典的方法，就是用已经标注情感类别和情感极性的词典来进行文本分析，这种词典即是情感词典。情感词典归纳了与情感相关的词汇，标注了每个词汇的情感类别和极性，还根据情感强烈程度进行了等级评定。分析时，需要提取文本中的每个词语，然后找到词语对应的情感类别和极性，分类汇总等级评分，就可以得出情感类别和极性各自的分数了。

小肖：原来是这样，那么基于机器学习的方法又是什么呢？

Miss 陈：基于机器学习的方法是指用机器学习算法，通过学习不同情感类别的文本，建立算法模型，然后用算法模型来识别新文章的情感类别。使用这类方法有个前提，就是必须事先收集大量的学习材料，即已经按情感类别分类的文章，作为建立模型的训练集，给算法建模。学习材料越多，模型效果越好。模型建好后，将新文章输入模型中计算情感类别。

比如要进行情感极性分析，需要收集尽可能多的“褒义”文章和“贬义”文章来建立模型，这实际上非常困难。如果能比较方便地获得分级文章，分析就轻松得多。比如豆瓣网的电影评论，每个评论都有对应的星级，总共五个星级。每个星级对应的评论就构成了这一等级的学习材料。根据这些材料进行机器学习，就能轻松建立算法模型，实现对新评论的自动分级。机器学习有不少算法，比如贝叶斯分类器、决策树、随机森林等，之前我们讲过的 Boosting 也属于机器学习算法。

小肖：这么说来用机器学习算法进行情感分析要复杂得多，主要是学习材料不好收集。

Miss 陈：是的。所以我们这次就选择相对容易、较好实现的基于情感词典的分析方法吧。

小肖：但是到哪里去找情感词典呢？

Miss 陈：关于情感词典，英文版的多，中文版的少。毕竟这方面的研究还是国外的起步早、研究多。不过随着我国研究的发展，也开发出了一些中文版情感词典，包括以下几种。

(1) 台湾大学研发的中文情感极性词典 NTUSD。

(2) 大连理工大学的情感本体词汇。

(3) 知网发布的“情感分析用词语集(beta 版)”。

(4) 哈尔滨工业大学社会计算与信息检索研究中心的《同义词词林》。

这些词典各有特色，且都免费提供使用。我们这次用的是大连理工大学的情感本体词汇。词典结构和部分内容见表 8-1。

表 8-1 情感本体词汇示例

词语	词性种类	词义数	词义序号	情感分类	强度	极性	辅助情感分类	强度	极性
脏乱	adj	1	1	NN	7	2			
糟报	adj	1	1	NN	5	2			
早衰	adj	1	1	NE	5	2			
责备	verb	1	1	NN	5	2			
贼眼	noun	1	1	NN	5	2			
战祸	noun	1	1	ND	5	2	NC	5	2
招灾	adj	1	1	NN	5	2			
折辱	noun	1	1	NE	5	2	NN	5	2
中山狼	noun	1	1	NN	5	2			
清峻	adj	1	1	PH	5				
清莹	adj	1	1	PH	5	1			
轻倩	adj	1	1	PH	5	1			
晴丽	adj	1	1	PH	5	1			

小肖：情感类别是怎么划分的呢？

Miss 陈：以大连理工大学的情感本体词汇为例，情感词汇被划分为七个类别，每个类别又细分为若干子类别，共有 20 个子类别。看看表 8-2 的内容你就明白了。

表 8-2　情感类别划分

情感大类	情感小类	例　　词
乐	快乐	喜悦、欢喜、笑眯眯、欢天喜地
	安心	踏实、宽心、定心丸、问心无愧
好	尊敬	恭敬、敬爱、毕恭毕敬、肃然起敬
	赞扬	英俊、优秀、通情达理、实事求是
	相信	信任、信赖、可靠、毋庸置疑、
	喜爱	倾慕、宝贝、一见钟情、爱不释手
怒	愤怒	气愤、恼火、大发雷霆、七窍生烟
哀	悲伤	忧伤、悲苦、心如刀割，悲痛欲绝
	失望	憾事、绝望、灰心丧气、心灰意冷
	疚	内疚、忏悔、过意不去、问心有愧
	思	相思、思念、牵肠挂肚、朝思暮想
惧	慌	慌张、心慌、不知所措、手忙脚乱
	恐	胆怯、害怕、担惊受怕、胆战心惊
	羞	害羞、害臊、面红耳赤、无地自容
恶	烦闷	憋闷、烦躁、心烦意乱、自寻烦恼
	憎恶	反感、可耻、恨之入骨、深恶痛绝
	贬责	呆板、虚荣、杂乱无章、心狠手辣
	妒忌	眼红、吃醋、醋坛子、嫉贤妒能
	怀疑	多心、生疑、将信将疑、疑神疑鬼
惊	惊奇	奇怪、奇迹、大吃一惊、瞠目结舌

小肖：明白了。大连理工大学的情感本体词汇把情感分为了七类，

每个子类别也很清晰。请问这个类别跟咱们所说的七情六欲中的七情有关系吗?

Miss 陈:有一定的对应关系,这方面你可以自行研究。

8.2.2 文本内容的分词方法

小肖:有了情感词典我们就可以进行分析了吧。不过,怎样才能把一篇文章的词语提取出来呢?我们的文章中每句话的词语都是连在一起的,不像英文那样是分开的,有空格间隔,不太好区分啊。

Miss 陈:很好,你发现了一个重要的问题。中文句子中的词语没有明显的间隔,要想区分并提取词语,难度要比英语大,这涉及另一个领域,即中文分词技术。

小肖:中文分词技术是什么意思呢?

Miss 陈:就是将中文句子分解成词语。“英文以空格作为天然的分隔符,而中文由于继承古代汉语的传统,词语之间没有分隔。古代汉语中除了联绵词和人名、地名等,词通常就是单个汉字,所以当时没有分词书写的必要。”而现代汉语中双字或多字词居多,一个字不再等同于一个词,但由于沿用古代汉语习惯,句中没有间隔区分词语。

要想从中文句子中分解出词语,需要用到分词算法。“现有的分词算法可分为三大类:基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。按照是否与词性标注过程相结合,又可以分为单纯分词方法和分词与标注相结合的一体化方法。”①

互联网搜索引擎就用到了分词技术。比如谷歌和百度搜索引擎,会根据输入的内容,通过分词算法提取词语,然后找到关键词,匹配搜索结果。

---

① 引用自必应网典。

小肖：原来中文句子分解成词语需要这么复杂的技术啊，那怎么办呢，我们有什么办法对句子进行分词的操作呢？

Miss 陈：一些高校有专业研究人员从事这方面的研究，并且向社会贡献出了他们的研究成果。我们这次就要用到中科院的 Ictclas 中文分词算法，采用隐马尔科夫模型（Hidden Markov Model，HMM）编写的 java 分词程序，在 R 语言中可以调用该算法进行分词。

小肖：要怎么做呢？

Miss 陈：其实做起来很简单，分词的速度很快，眨眼工夫就可以完成一篇文章的分词，来试试吧！

8.3 分析过程

8.3.1 导入分析内容

Miss 陈：首先，我们需要在 R 语言中导入分析材料。这就像做菜，先将食材准备好，然后才开始烹饪。要导入的材料包括辞职报告文本和情感本体词库。

小肖：这些材料都有了。辞职报告是 Word 版本，情感本体词库是 Excel 版本，可以直接用吗？

Miss 陈：需要转换一下。辞职报告要转换为纯文本，后缀名是 txt.；情感本体词库要转换为后缀名是 csv. 格式的版本。

小肖：好的，我马上转换一下。嗯，转换好了。

Miss 陈：那么现在可以导入这些内容了。

导入辞职报告和情感本体词库的 R 语句如下：

```
#导入辞职报告
myfile=scan("第八章/辞职报告模板.txt", what="", sep="\n")
#导入情感本体词库
mydict<-read.csv("第八章/情感词汇本体.csv")
#获得褒义词库
mydict.p.Word<-subset(mydict,mydict$极性=="1")
#获得贬义词库
mydict.n.Word<-subset(mydict,mydict$极性=="2")
#获得中性词库
mydict.m.Word<-subset(mydict,mydict$极性=="0")
#获得词库：乐
mydict.le<-subset(mydict,mydict$情感分类==c("PA","PE"))
#获得词库：好
mydict.ha<-subset(mydict,mydict$情感分类==c("PD","PH","PG",
"PB","PK"))
#获得词库：怒
mydict.lu<-subset(mydict,mydict$情感分类=="NA")
#获得词库：哀
mydict.ai<-subset(mydict,mydict$情感分类==c("NB","NJ","NH",
"PF"))
#获得词库：惧
mydict.ju<-subset(mydict,mydict$情感分类==c("NI","NC","NG"))
#获得词库：恶
mydict.wu<-subset(mydict,mydict$情感分类==c("NE","ND","NN",
"NK","NL"))
#获得词库：惊
mydict.ji<-subset(mydict,mydict$情感分类=="PC")
```

8.3.2 分词

Miss 陈：接下来进行分词操作，把导入的辞职报告中的词语提取出来。

小肖：是调用中国科学院的 Ictclas 中文分词算法进行分词吧？

Miss 陈：是的。正常情况下，分词需要过滤一些内容，包括以下两方面。

(1) 过滤标点符号和空格,因为这些不是词语。

(2) 过滤停用词。停用词是指没有实际含义或分析价值低的词语,包括英文字符、数字、数学字符、使用频率高的单个汉字等。但进行文本情感分析不用过滤停用词,因为运用情感词典进行分析时会自动过滤掉这些词语。这篇辞职报告分词后结果如下:

“尊敬”、“的”、“领导”、“在”、“递交”、“这”、“份”、“辞呈”、“时”、“我”、“的”、“心情”、“十分”、“矛盾”、“现在”、“公司”、“的”、“发展”、“需要”、“大家”、“竭尽全力”、“由于”、“我”、“状态”、“不”、“佳”、“和”、“一些”、“个人”、“原因”、“的”、“影响”、“无法”、“为”、“公司”、“做出”、“相应”、“的”、“贡献”、“因此”、“请求”、“允许”、“离开”、“从”、“昨天”、“晚上”、“到”、“今天”、“是”、“继续”、“坚持”、“还”、“是”、“果断”、“放弃”、“这个”、“问题”、“一直”、“困扰”、“着”、“我”、“经过”、“一天”、“一夜”、“的”、“考虑”、“我”、“还”、“是”、“选择”、“放弃”、“曾经”、“那么”、“大”、“的”、“风浪”、“都”、“挺”、“过来”、“了”、“多么”、“难”、“过”、“的”、“坎”、“也”、“过去”、“了”、“如今”、“为什么”、“就”、“选择”、“放弃”、“呢”、“这个”、“问题”、“我”、“也”、“问”、“过”、“自己”、“回想”、“这”、“两年”、“来”、“其实”、“我”、“比”、“任何人”、“都”、“珍惜”、“这”、“份”、“工作”、“我”、“知道”、“我”、“不”、“是”、“最”、“优秀”、“的”、“但”、“我”、“是”、“非常”、“努力”、“的”、“值得”、“庆幸”、“的”、“是”、“至少”、“学生”、“对”、“我”、“的”、“责任心”、“和”、“努力”、“付出”、“是”、“非常”、“肯定”、“的”、“仅”、“凭”、“学生”、“的”、“几句”、“话”、“我”、“就”、“知道”、“我”、“这”、“两年”、“的”、“努力”、“是”、“值得”、“的”、“企业”、“和”、“领导”、“对”、“员工”、“的”、“肯定”、“与”、“鼓励”、“关心”、“与”、“爱护”、“就”、“如同”、“老师”、“对”、“学生”、“的”、“肯定”、“与”、“鼓励”、“关心”、“与”、“爱护”、“让”、“学生”、“在”、“精神”、“上”、“取得”、“莫大”、“的”、“鼓舞”、“从而”、“会”、“更加”、“自觉”、“地”、“努力”、“学习”、“对于”、“员工”、“来说”、“也”、“会”、“发

自”、“内心”、“的”、“心甘情愿”、“地”、“为”、“企业”、“做”、“贡献”、“使”、“企业”、“得以”、“更”、“好”、“的”、“发展”、“其实”、“我”、“认为”、“这”、“是”、“相互”、“的”、“老师”、“对”、“学生”、“不”、“付出”、“爱”、“学生”、“当然”、“也”、“不”、“会”、“爱”、“老师”、“少”、“一点”、“私心”、“多一点”、“无私”、“真心”、“真意”、“为”、“学生”、“着想”、“学生”、“才”、“会”、“真的”、“爱”、“老师”、“班级”、“才”、“会”、“团结”、“稳定”、“从”、“2006 年”、“9 月”、“带”、“A 班”、“和”、“B 班”、“开始”、“我”、“一直”、“这么”、“认为”、“也”、“一直”、“这样”、“努力”、“着”、“可是”、“渐渐”、“地”、“我”、“发现”、“因为”、“各”、“方面”、“的”、“因素”、“这”、“一点”、“越来越”、“难”、“做到”、“了”、“可能”、“是”、“我”、“太”、“单纯”、“可能”、“是”、“我”、“不”、“成熟”、“可能”、“是”、“我”、“太”、“理想化”、“我”、“感到”、“迷茫”、“和”、“困惑”、“做”、“任何”、“事”、“我”、“都”、“很”、“努力”、“很”、“认真”、“我”、“不”、“想”、“混”、“那样”、“会”、“让”、“我”、“良心”、“不安”、“同时”、“我”、“胆小”、“做事”、“之前”、“都”、“要”、“思来想去”、“瞻前顾后”、“害怕”、“出现”、“自己”、“不能”、“预料”、“的”、“后果”、“缺乏”、“魄力”、“这”、“也”、“是”、“我”、“在”、“工作”、“上”、“难以”、“得到”、“发展”、“的”、“致命”、“缺”、“点”、“不”、“果断”、“是”、“我”、“另”、“一个”、“致命”、“的”、“缺”、“点”、“如今”、“我”、“的”、“工作”、“也”、“真”、“的”、“走”、“到”、“了”、“瓶颈”、“处”、“因此”、“我”、“不得不”、“离开”、“也许”、“此时”、“提出”、“辞呈”、“会”、“显得”、“不”、“合适”、“公司”、“正”、“处于”、“快速”、“发展”、“的”、“阶段”、“同事”、“们”、“都”、“是”、“斗志昂扬”、“壮志”、“满怀”、“而”、“我”、“在”、“这时候”、“却”、“因”、“个人”、“原因”、“无法”、“为”、“公司”、“分忧”、“实在”、“是”、“深感”、“歉意”、“其实”、“我”、“也”、“很”、“不”、“舍”、“舍不得”、“那些”、“既”、“让”、“人”、“爱”、“又”、“让”、“人”、“恨”、“的”、“学生”、“舍不得”、“相处”、“了”、“两年”、“的”、“同事”、“感谢”、“向”、“老师”、“一直”、“以来”、“对”、“我”、

“的”、“关心”、“和”、“教导”、“也”、“感谢”、“肖”、“老师”、“平时”、“对”、“我”、“的”、“关心”、“和”、“帮助”、“在”、“他们”、“身上”、“有”、“很多”、“值得”、“我”、“学习”、“的”、“地方”、“向”、“老师”、“超”、“强”、“的”、“处事”、“能力”、“肖”、“老师”、“圆滑”、“的”、“为人”、“处世”、“风格”、“这些”、“都”、“是”、“我”、“非常”、“缺乏”、“的”、“更加”、“感谢”、“公司”、“给”、“了”、“我”、“做”、“老师”、“的”、“机会”、“感谢”、“公司”、“所有”、“领导”、“和”、“同事”、“对”、“我”、“的”、“教诲”、“和”、“关心”、“这”、“两年”、“我”、“也”、“收获”、“颇”、“丰”、“最后”、“我”、“有”、“一个”、“请求”、“从”、“去年”、“周年”、“庆”、“到”、“今年”、“周年”、“庆”、“我”、“又”、“工作”、“了”、“整整”、“一年”、“虽然”、“没有”、“取得”、“过”、“人”、“的”、“成绩”、“但”、“我”、“是”、“在”、“勤勤恳恳”、“地”、“做事”、“我”、“从来不”、“会”、“提出”、“要求”、“这”、“也”、“是”、“我”、“的”、“第三个”、“缺点”、“这”、“是”、“我”、“第一次”、“也”、“是”、“最后”、“一次”、“要求”、“恳请”、“领导”、“在”、“结算”、“工资”、“时”、“将”、“周年”、“庆”、“奖金”、“和”、“2006 年”、“9 月”、“2007 年”、“12 月”、“所”、“欠”、“的”、“社保”、“费用”、“连同”、“工资”、“一起”、“结算”、“给”、“我”、“不”、“胜”、“感激”、“我”、“希望”、“公司”、“领导”、“在”、“百”、“忙”、“之中”、“抽出”、“时间”、“商量”、“一下”、“工作”、“交接”、“问题”、“本人”、“将于”、“2008 年”、“9 月”、“5 日”、“离职”、“希望”、“得到”、“领导”、“的”、“准许”、“感谢”、“诸位”、“在”、“我”、“在”、“公司”、“期间”、“给予”、“我”、“的”、“信任”、“和”、“支持”、“并”、“祝”、“所有”、“同事”、“和”、“朋友”、“们”、“在”、“工作”、“和”、“活动”、“中”、“取得”、“更”、“大”、“的”、“成绩”、“和”、“收益”、“此致”、“敬礼”

小肖：分词的效果不错啊，速度也很快，果然眨眼之间就出结果了。

Miss 陈：是的。分词速度能够如此之快，和现在中文文本分析技术

的发展有很大的关系。中文分词算法发展了好几代，由于 R 语言的普及，有爱好者将中文分词的功能整合到了其中，于是在 R 语言中进行中文分词得以实现。早几年想进行中文分词可不是件容易的事情。这次我们只是计算情感积分，所以分词后不需要对文本进行其他加工。

小肖：什么是其他加工呢？

Miss 陈：就是进一步处理分词后的结果，包括去掉停用词、判断词性、建立语料库等，以后碰到这些情况时再讨论。

分词的 R 语句如下：

```
#分句,去标点
myfile<-strsplit(myfile,split="")
myfile.split<-unlist(myfile)
#去空格
myfile.split<-str_trim(myfile.split)
#分词
myfile.Words<-lapply(myfile.split,FUN=segmentCN)
myfile.Words<-as.vector(myfile.Words)
```

8.3.3 计算情感积分

Miss 陈：然后就是最重要的工作：计算各类情感的积分。

小肖：怎么计算呢？

Miss 陈：分词后计算情感积分就比较简单了，依次根据情感词库中对应的词语，找到该词语所属情感类别的情感强度评分，将各类情感分数累加起来，就得到了每类情感的积分。

小肖：明白了，这步操作有点像 Excel 中的条件查询和分类汇总。

Miss 陈：是的。

情感类别积分的结果见表 8-3。

表 8-3　情感类别积分计算结果

序号	类别	积　分
1	乐	0.196 261 68
2	好	0.654 205 61
3	怒	0.000 000 00
4	哀	0.093 457 94
5	惧	0.028 037 38
6	恶	0.028 037 38
7	惊	0.000 000 00

情感极性积分的结果见表 8-4。

表 8-4　情感极性积分计算结果

序号	类别	积　分
1	褒义	0.489 913 5
2	中性	0.391 930 8
3	贬义	0.118 155 6

计算情感类别积分的 R 语句如下：

```
#匹配词库：乐
fileScore.le=0
for(i in 1:length(myfile.Words)){
  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)
  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.le$强度, mydict.le$词语==myfile.sentence.
    Word)
if(length(x)>0){senScore=senScore+x}
}
  print(i/length(myfile.Words))
```

```
if(length(senScore)>0){fileScore.le=fileScore.le+senScore }
}
#匹配词库：好
fileScore.ha=0
for(i in 1:length(myfile.Words)){
  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)
  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.ha$强度, mydict.ha$词语==myfile.sentence.
    Word)
if(length(x)>0){senScore=senScore+x}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.ha=fileScore.ha+senScore}
}
#匹配词库：怒
fileScore.lu=0
for(i in 1:length(myfile.Words)){
  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)
  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.lu$强度, mydict.lu$词语==myfile.sentence.
    Word)
if(length(x)>0){senScore=senScore+x}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.lu=fileScore.lu+senScore}
}
#匹配词库：哀
fileScore.ai=0
for(i in 1:length(myfile.Words)){
  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)
```

```
  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.ai$强度, mydict.ai$词语==myfile.sentence.
    Word)
if(length(x)>0){senScore=senScore+x}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.ai=fileScore.ai+senScore}
}
#匹配词库：惧
fileScore.ju=0
for(i in 1:length(myfile.Words)){
  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)
  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.ju$强度, mydict.ju$词语==myfile.sentence.
    Word)
if(length(x)>0){senScore=senScore+x}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.ju=fileScore.ju+senScore}
}
#匹配词库：恶
fileScore.wu=0
for(i in 1:length(myfile.Words)){
  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)
  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.wu$强度, mydict.wu$词语==myfile.sentence.
    Word)
if(length(x)>0){senScore=senScore+x}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.wu=fileScore.wu+senScore}
```

```
}
#匹配词库：惊
fileScore.ji=0
for(i in 1:length(myfile.Words)){
  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)
  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.ji$强度, mydict.ji$词语==myfile.sentence.
    Word)
if(length(x)>0){senScore=senScore+x}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.ji=fileScore.ji+senScore}
}
#计算情感类别积分
fileScore.le.pert<-fileScore.le/(fileScore.le+fileScore.ha+
fileScore.lu+fileScore.ai+fileScore.ju+fileScore.wu+fileScore.
ji)
fileScore.ha.pert<-fileScore.ha/(fileScore.le+fileScore.ha+
fileScore.lu+fileScore.ai+fileScore.ju+fileScore.wu+fileScore.
ji)
fileScore.lu.pert<-fileScore.lu/(fileScore.le+fileScore.ha+
fileScore.lu+fileScore.ai+fileScore.ju+fileScore.wu+fileScore.
ji)
fileScore.ai.pert<-fileScore.ai/(fileScore.le+fileScore.ha+
fileScore.lu+fileScore.ai+fileScore.ju+fileScore.wu+fileScore.
ji)
fileScore.ju.pert<-fileScore.ju/(fileScore.le+fileScore.ha+
fileScore.lu+fileScore.ai+fileScore.ju+fileScore.wu+fileScore.
ji)
fileScore.wu.pert<-fileScore.wu/(fileScore.le+fileScore.ha+
fileScore.lu+fileScore.ai+fileScore.ju+fileScore.wu+fileScore.
ji)
fileScore.ji.pert<-fileScore.ji/(fileScore.le+fileScore.ha+
fileScore.lu+fileScore.ai+fileScore.ju+fileScore.wu+fileScore.
ji)
```

计算情感极性积分的 R 语句如下：

```
#匹配褒义词库
fileScore.P=0
for(i in 1:length(myfile.Words)){

  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)

  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.p.Word$强度, mydict.p.Word$词语==myfile.
    sentence.Word)
if(length(x)>0){senScore=senScore+x[1]}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.P=fileScore.P+senScore}
}
#匹配贬义词库
fileScore.N=0
for(i in 1:length(myfile.Words)){

  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)



  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.n.Word$强度, mydict.n.Word$词语==myfile.
    sentence.Word)
if(length(x)>0){senScore=senScore+x[1]}
}
  print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.N=fileScore.N+senScore}
}
```

```
#匹配中性词库
fileScore.M=0
for(i in 1:length(myfile.Words)){

  myfile.sentence.Word<-unlist(myfile.Words)
  myfile.sentence.Word<-as.list(myfile.sentence.Word)
  myfile.sentence.Word<-as.vector(myfile.sentence.Word)

  senScore=0
for(j in 1:length(myfile.sentence.Word)){
    x<-subset(mydict.m.Word$强度, mydict.m.Word$词语==myfile.
    sentence.Word)
  if(length(x)>0){senScore=senScore+x[1]}
}
print(i/length(myfile.Words))
if(length(senScore)>0){fileScore.M=fileScore.M+senScore}
}

#计算情感极性积分
fileScore=0
fileScore=fileScore.P - fileScore.N
print(fileScore)
#计算比例
fileScore.P.pert<-fileScore.P/(fileScore.P+fileScore.N+
fileScore.M)
fileScore.M.pert<-fileScore.M/(fileScore.P+fileScore.N+
fileScore.M)
fileScore.N.pert<-fileScore.N/(fileScore.P+fileScore.N+
fileScore.M)
```

8.3.4 显示结果

小肖：计算情感类别积分和情感极性积分的 R 语句很长啊。

Miss 陈：虽然代码长，其实有规律可循。你仔细看可以发现，计算每一类情感的语句基本都是一样的。

小肖：那我得睁大眼睛仔细看看。

Miss 陈：计算完成后，就可以根据分数计算结果绘制条形图，让情感分析的结果更加直观。辞职报告的情感类别分析结果如图 8-3 所示。

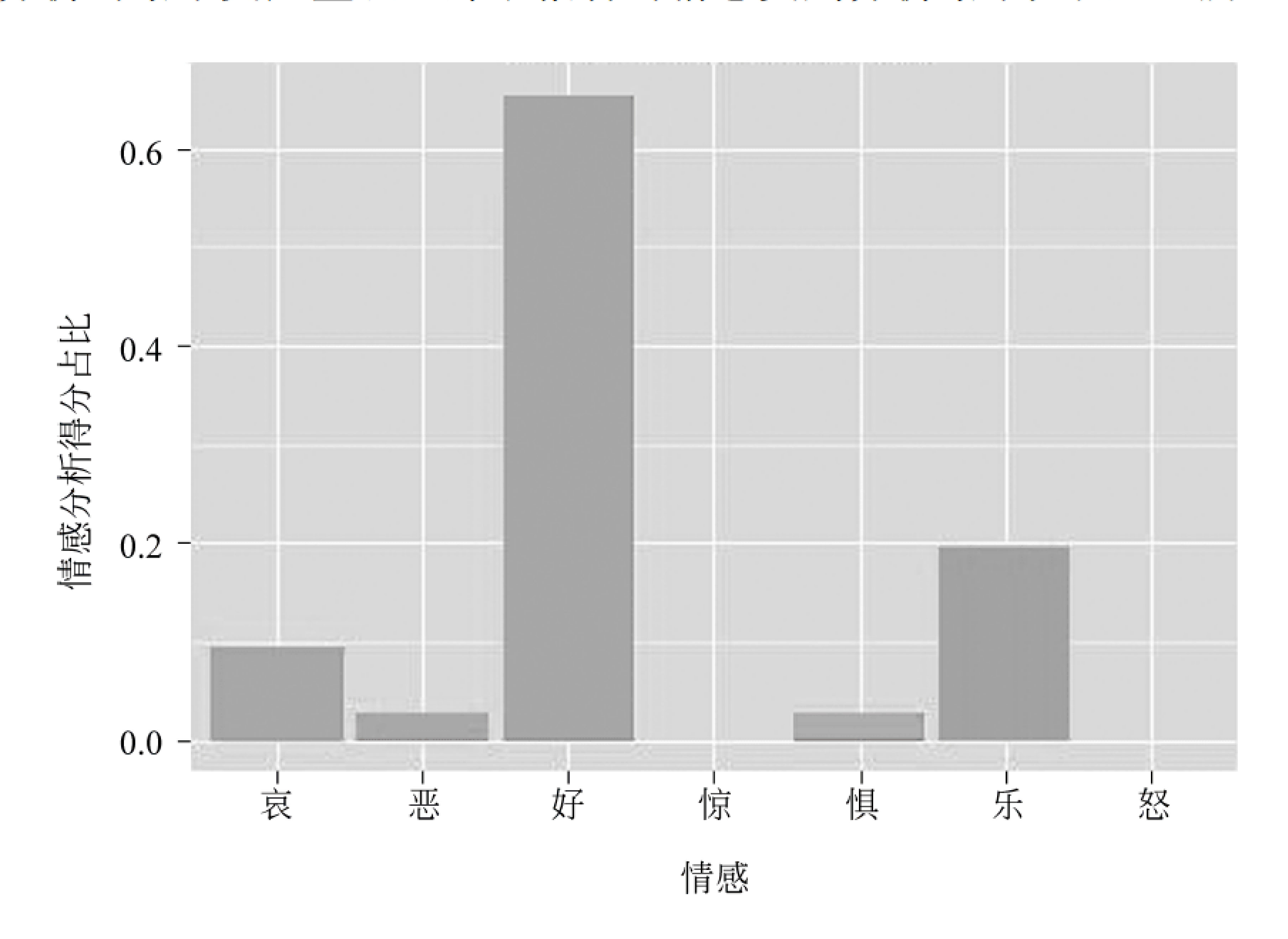


图 8-3　辞职报告的情感类别分析结果

辞职报告的情感类别分析结果绘图的 R 语句如下：

```
#形成数据框
fileScore.frame<-data.frame(name=c("乐","好","怒","哀","惧",
"恶","惊"), Score = c (fileScore.le.pert, fileScore.ha.pert,
fileScore.lu.pert, fileScore.ai.pert, fileScore.ju.pert,
fileScore.wu.pert,fileScore.ji.pert))
#绘图
qplot(x=name,y=Score,data=fileScore.frame,geom="bar", stat="
identity",xlab="情感",ylab="情感分析得分占比",fill=name,main="辞
职报告情感类别分析结果")
```

辞职报告的情感极性分析结果如图 8-4 所示。

辞职报告的情感极性分析结果绘图的 R 语句如下：

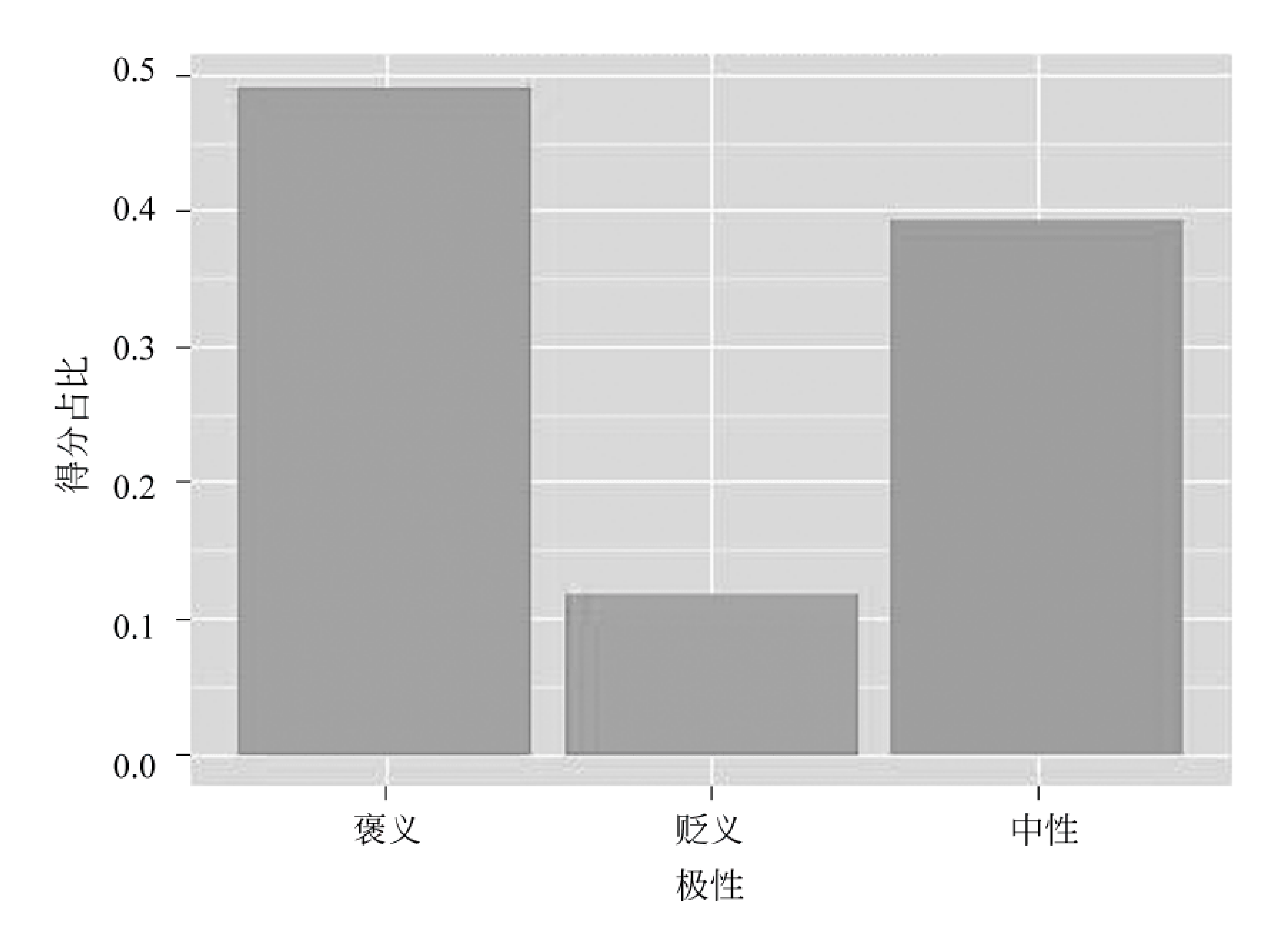


图 8-4　辞职报告的情感极性分析结果

```
#形成数据框
fileScore.frame<-data.frame(name=c("褒义","中性","贬义"),Score
=c(fileScore.P.pert,fileScore.M.pert,fileScore.N.pert))
#绘图
qplot(x=name,y=Score,data=fileScore.frame,geom="bar", stat=
"identity",xlab="极性",ylab="得分占比",fill=name,main="辞职报告
情感极性分析结果")
```

到这步辞职报告的情感分析就算完成了，分析结果已可以在实际管理中应用。

小肖：这种文本量化分析技术和其他的数据分析技术不大一样，真是让人大开眼界，没想到辞职报告还可以这样分析，一篇文章也能进行量化。

Miss 陈：文本分析技术是数据分析的一个另类领域，实际上用途广泛。比如，网络舆情分析（根据微博、微信内容分析舆论倾向）、文章内容自动推荐（根据用户喜好建立预测模型，主动推送用户感兴趣的文章）等。

反而对于我们人力资源管理专业领域的应用还不多，这方面需要我们去探索和发现。

人力资源管理数据分析就谈到这里吧。我希望你们能明白，在统计学领域有许多知识、算法、工具可以应用到我们人力资源管理的实践中，但前提是我们要主动学习这方面的知识，懂这方面的技术。我希望以后能看到更多数据分析技术应用到人力资源管理中来，促进我们管理水平的提升。

小肖：明白了，谢谢经理！我们一定努力学习数据分析知识，掌握相关分析工具，特别是重点学习 R 语言，掌握这个强大的分析工具。